UAB
Universitat Autònoma de Barcelona

Morphometric, histological, and genomic study of the natural variability of apple size and shape

Eduardo Dujak Riquelme

Doctoral thesis

Universitat Autònoma de Barcelona
Facultat de Biociències
Departament de Biologia Animal, Biologia Vegetal i Ecologia

Morphometric, histological, and genomic study of the natural variability of apple size and shape

Dissertation presented by Christian Eduardo Dujak Riquelme for the degree of Doctor in Plant Biology and Biotechnology by Universitat Autònoma de Barcelona

Thesis director and tutor

Dr. Maria José Aranzana Civit

PhD candidate

Christian Eduardo Dujak Riquelme

Barcelona, December 2022

This thesis has been performed in the research group of Rosaceae Genetics and Genomics, from the research program of Plant and Animal Genomics at the Centre for Research in Agricultural Genomics (CRAG) - CSIC-IRTA-UAB-UB consortium. The PhD student, Christian Eduardo Dujak Riquelme, has been supported by "DON CARLOS ANTONIO LOPEZ" Abroad Postgraduate Scholarship Program, BECAL-Paraguay, which also funded an international stage at the Plant and Food Research (P&FR), at the Molecular and Digital Breeding Division (Palmerston North, New Zealand).

Ojededikava che sype

Acknowledgments

Quiero empezar estas lineas con unas palabras de un ilustre que dice,

"La raíz de todo bien crece en la tierra de la gratitud"

Como empezar a dar gracias por años de esfuerzo y dedicacion pura a la ciencia. Quisiera dar gracias primeramente a mi directora Dra. Maria Jose Aranzana, por haber confiado en mi y abrir las puertas a su grupo de trabajo, por formarme con criterio cientifico y ayudarme a tomar las desiciones dia a dia para esta tesis. Tambien quiero agradecer a Ibo por las ayudas continuas desde el vamos con el mapeo genetico hasta en los dias de campo, siempre con una sonrisa y amable en poder ayudarte. A los pilares del laboratorio, Elena (mi madre del laboratorio), Fuensi y Angel sin ellos cada uno de los estudiantes que pasaron por el laboratorio sabemos que seria imposible solucionar solos algunos protocolos de laboratorio que no van por algun motivo, pero con la experiencia de tantos años te dan refugio para continuar el arduo trabajo, y sobre todo por tener la amabiliadad, las ganas de enseñar y el humor positivo dia a dia.

Sabemos que durante los años de tesis, hemos compartido con compañeros de oficina o del grupo de trabajo. Por ello quiero agradecer al Dr. Fiol porque desde que llegue al CRAG, fue el primero en darme una mano en todo lo va relacionado al uso de herramientas de laboratorio, al pasar el tiempo se hizo como un hermano mayor que me enseñaba como tener equilibrio en estado etilitico miestras hacias extraccion de RNA, por las tardes sangrientas de Virus con mi compipa Carla aquien aprecio un monton. Espero en un futuro cercano que podamos trabjar en conjunto. Tambien a Nuria, Neus, Nathalia, Miguel, Pol, Carlos, Federico por darme una mano en disipar mis dudas con respecto a mis experimentos y por sobre todo la buena compañía el dia a dia. También agradecer a los estudiantes que han pasado por el proyecto: Laura, Alejandro y Laia. Espero que la experiencia le sirviera tanto como me ha servido a mí. A Los compañeros del IRTA, que cada miercoles de evaluacion de campo estuvieron para darte una mano. Y por ultimo a mi familia y compañero de lucha que en realidad no podria seguir sin sus ayudas, animos continuos en dias grises o claros. Son mi pilar el dia a dia. A todos: muchas gracias por ser el apoyo.

Index of contents

Index of figures and tables

CHAPTER 3

Figures

Abbreviations

AFLP: Amplified fragment length polymorphism

BLAST: Basic Local Alignment Search Tool

bp: Base pair

Chr: Chromosome

CRISPR: Clustered regularly interspaced short palindromic repeat

DNA: Deoxyribonucleic acid

ECPGR: European Cooperative Programme for Plant Genetic Resources

GBS: genotyping by sequencing

GDR: Genome Database for Rosaceae

GDDH13: Golden Delicious doble haploide 13

GWAS: Genome wide association studies

HiDRAS: High-quality disease resistant apples for a sustainable agriculture

IGV: Integrative Genomics Viewer

Kb: Kilobase

LD: Linkage desequilibrium

LG: Linkage group

LOD: Logarithm of the odds

MAS: Marker-assisted selection

Mb: Megabase

mRNA: Messenger RNA

NCBI: National Center for Biotechnology Information

OECD: Organisation for Economic Co-operation and Development

PCR: Polymerase chain reaction

qPCR: quantitative PCR

QTL: Quantitative trait locus

QTLs: Quantitative trait loci

QTN: Quantitative trait nucleotide

RAPD: Randomly amplified polymorphic DNA

RFLP: Restriction fragment length polymorphism

RNA: Ribonucleic acid

RNA-seq: RNA sequencing

RT: retrotranscriptase

RT-PCR: Reverse transcription PCR

SNP: Single nucleotide polymorphism

SSR: Simple sequence repeat

SV: Structural variant

ABSTRACT

ABSTRACT

Apple size and shape are sensory traits that influence in the consumers purchase decisions. While breeders breed for “nice” shape apples, cultivar characterization requires a more precise description of the shape concept. Variety and cultivar evaluators use agreed guidelines and sketches to assign apples into classes. Regardless of the allowable margin of error due to subjectivity in the assignments, these descriptors are of less use for genetic studies. In this work, we did a morphometric analysis of apple fruits evaluating ~13,000 2D images of sections of 364 genotypes of the Apple REFPOP, using Tomato Analyzer software. Data analysis allowed for an in-depth characterization of apple morphology. A Random Forest analysis stablished FSII (ration between height and width) as the most relevant trait determining the fruit shape, followed by the FST (indicator of conicity). Morphometric data were used in two GWAS models (FarmCPU and BLINK) that found 59 SNPs associated with fruit size and shape traits. The haploblocks containing the most relevant SNPs served to propose candidate genes. Histological evaluations of fruits of three cultivars with contrasting shape (flat, round and oblate) at 0, 61 and 98 days after anthesis (DAA). RNA-seq data served to identify differentially expressed genes (DEG) along development (TC-DEA) and between cultivars. Some were phytohormones with a role in fruit development. GWAS and TC-DEA analysis identified the gene *MdOFP4* as a strong candidate for fruit shape. A polymorphism in the promoter of the gene could be the reason of its null expression in the cultivar SKO, producing the oblate shape.

RESUMEN

El tamaño y la forma de la manzana son características sensoriales que influyen en las decisiones de los consumidores. Mientras que los mejoradores buscan manzanas de forma “bonita”, la caracterización del cultivares requiere una descripción más precisa del concepto de forma. Los examinadores de variedades y germoplasma utilizan pautas y esquemas acordados para asignar manzanas a clases. Independientemente del margen de error asumible debido a la subjetividad en las asignaciones, estos descriptores son de menor utilidad para estudios genéticos. En este trabajo, hicimos un análisis morfométrico de frutos de manzana evaluando ~13,000 imágenes 2D de secciones de 364 genotipos de AppleREFPOP, utilizando el software Tomato Analyzer. El análisis de datos permitió una caracterización en profundidad de la morfología de la manzana. Un análisis de Random Forest estableció FSII como el parámetro más relevante para determinar la forma de la fruta, seguido por el FST. Los datos morfométricos se utilizaron en dos modelos GWAS (FarmCPU y BLINK) que identificaron 59 SNP asociados con el tamaño y la forma de la manzana. La construcción de haplobloques en los los SNP más relevantes sirvieron para proponer genes candidatos. Se realizaron evaluaciones histológicas en frutos de tres cultivares con diferente forma (plano, redondo y oblongo) a los 0, 61 y 98 días después de la antesis (DAA). Los datos de RNA-seq sirvieron para identificar genes expresados diferencialmente (DEG) a lo largo del desarrollo (TC-DEA) y entre cultivares. Algunos eran fitohormonas relevantes para el desarrollo del fruto. El análisis GWAS y TC-DEA identificó el gen *MdOFP4* como un fuerte candidato para la forma de la fruta. Un polimorfismo en el promotor del gen podría ser la causa de que no se exprese en el cultivar SKO, produciendo la forma oblonga.

RESUM

La mida i la forma de la poma són característiques sensorials que influeixen en les decisions dels consumidors. Mentre que els milloradors busquen pomes que tinguin forma “bonica”, la caracterització del cultivars requereix una descripció més precisa del concepte de forma. Els examinadors de varietats i germoplasma utilitzen pautes i esquemes acordats per assignar pomes en classes. Independentment del marge d’error acceptable a causa de la subjectivitat en les assignacions, aquests descriptors són de menor utilitat per a estudis genètics. En aquest treball, vam fer una anàlisi morfomètrica de pomes avaluant ~13,000 imatges 2D de seccions de 364 genotips d' Apple REFPOP, utilitzant el program Tomato Analyzer (TA). L'anàlisi de dades va permetre una caracterització en profunditat de la morfologia de la poma. Una anàlisi de Random Forest va establir FSII com el paràmetre més rellevant per determinar la forma de la fruita, seguit pel FST. Les dades morfomètriques es van utilitzar en dos models GWAS (FarmCPU i BLINK) que van identificar 59 SNP associats amb la mida i la forma de la poma. La construcció d'haploblocs als SNP més rellevants van servir per proposar gens candidats. Es van realitzar avaluacions histològiques en fruits de tres cultivars amb diferent forma (plànol, rodó i oblong) als 0, 61 i 98 dies després de l'antesi (DAA). Les dades de RNA-seq van servir per identificar gens expressats diferencialment (DEG) al llarg del desenvolupament (TC-DEA) i entre cultivars. Alguns eren fitohormones rellevants per al desenvolupament del fruit. L'anàlisi GWAS i TC-DEA va identificar el gen *MdOFP4* com a fort candidat per a la forma de la fruita. Un polimorfisme en el promotor del gen podria ser la causa de que no s’expressi en el cultivar SKO, produint la forma oblonga.

MAIN INTRODUCTION

SECTION 1: APPLE GENERAL ASPECTS

1.1 Taxonomy

Apple belongs to the Malus genus within the Rosaceae family. The cultivated apple is generally denominated *Malus* × *domestica* Borkh, while the scientific name *Malus pumilla* Mill. is also accepted among the scientific community (Korban and Skirvin, 1984). This species is taxonomically classified as follows:

Kingdom: Plantae

Plylum: Tracheophyta

Class: Magnoliopsida

Order: Rosales

Family: Rosaceae

The Rosaceae family members are dicotyledonous and include most of the consumed fruit species: apple, pear, peach, plum, cherry, strawberry, almond, apricot, blackberry, crab apple. This family also includes many ornamental species as the roses. The Rosaceae family is subdivided into three subfamilies: Rosoideae, Dryadoideae and Amygdaloideae (Morgan et al., 1994; Xiang et al., 2017).

Subfamily: Amygdaloideae

Tribe: Maleae

The tribe Maleae is divided into eight Sections: *Sorbomalus*, *Yunnanenses*, *Sorbomalus*, *Malus*, *Gymnomeles*, *Chloromeles*, *Docyniopsis* and *Eriolobus*, and comprises around 29 to 31 wild species (Robinson et al., 2001; Qian et al., 2006; Xiang et al., 2017).

Genus: Malus Mill.

Species: *Malus* × *domestica* Borkh

1.2 Botanical characteristics

Apple trees grow in temperate regions. They are deciduous or semi-deciduous with alternate leaves, which are usually serrated and oval to ovate in shape (Pratt, 1993; Fischer, 1q994). In natural conditions, the trees can reach a height from 2 to 20 meters and take from 5 to 12 years to overcome the juvenile period.

During the first years following seed germination, the tree is orthotropic (i.e., grows in an erect axes) with monopodial lateral branching with rhythmic and indeterminate growth. Later, at the adult stage, the tree can develop sympodial branching when long shoots produce terminal flowers (Lauri and Laurens, 2005).

In branches, the buds can be mixed or vegetative. The mixed ones can develop apical or lateral and contain vegetative and reproductive primordia that will become into a spur, which consists of a short shoot (bourse) on which the leaf primordia will extend, as well as one or two shoots (bourse shoots) and the inflorescence. The vegetative bud develops into vegetative shoots.

Some characteristics of the fruiting branches are relevant for the fruit production. For example, the length and volume of their terminal bourse is related to the biennial yield (Lespinasse and Delort, 1993), probably because of its role in producing and transporting flower formation signals (Elsysy and Hirst, 2017).

In commercial apple production, crop load must be managed to maximize economic return. Therefore, the tree canopy is trained into shapes that improve yield and fruit quality with a fast entrance into production and efficient management, pruning and harvesting. Excessively light or heavy crop loads reduce fruit quality and may result in fruit sizes that have lower consumer acceptance, with the corresponding economic impact. The optimum system is cultivar dependent and should take into consideration

environmental aspects, as the climate, soil, and prevalence of certain pests (Lauri et al., 2016). In addition, the rootstock-cultivar-soil combination will have an impact on the tree response to a certain training system (Forshey et al., 1992).

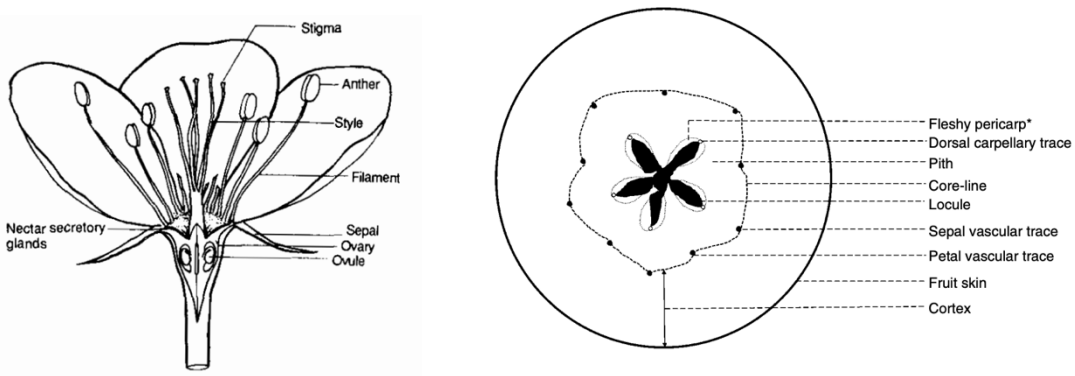
Root system

To facilitate the clonal propagation of apple genotypes, vegetative buds are usually grafted in rootstocks, which are usually developed in breeding programs with improved agronomic traits to the scion, as could be resistance to soil pathogens, to drought stress or to control the tree vigor (Marini and Fazio, 2018). Since there is an interchange of nutrients and hormones between the rootstock and the variety (scion), the correct selection of the rootstock is fundamental for an optimum establishment of the crop in a given edaphoclimatic environment. For instance, the dwarfing rootstock has become an important resource to control vigor while leaving the tree energy for the fruit-bearing, as well as reducing labor costs (Costes and Garcia-Villanueva, 2007; Fazio, 2021). In addition, the rootstock may induce graft precocity and regulate flowering intensity (Albacete et al., 2017).

Fruit and flower morphology

Inflorescences usually appear in the third or fourth year of the tree. The development of the floral primordia is initiated in summer and culminates in spring with the final formation of the floral organs (Koutinas et al., 2010). Inflorescences are in general conformed by five flowers at the base of the lateral vegetative buds (Eccher et al., 2014). The flowers are hermaphroditic and epigynous, with five petals from white to red color, a calyx of five sepals and around twenty stamens. Each stamen has a filament and an anther with two pollen sacs.

The flower has a unique ovary with five fused carpels (i.e., syncarpus ovary) with two ovules per carpel (**Figure 1A**). The ovary carpels are surrounded by non-ovarian tissue that will develop into a pseudocarpic fruit, called "pome fruit" (Janick et al., 1996; Dennis et al., 2003).



A **B**

Figure I.1. Parts of the flower and fruit of the apple tree**. A**, the apple flower drawn in longitudinal section, showing its parts (image was extracted from MacDaniels and Heinicke (1929). **B**, apple fruit in diagrammatic transversal section, showing the primary tissues [image from Malladi (2020)].

From the outside to the inside of the fruit, the "poma" or apple is composed by primary skin, which contains cellular and polymeric compounds, and is structured into the cuticle, the epidermis and the hypodermis. The cuticle is the outermost layer, made of cutin and wax that deposit in the outer cell wall of the epidermis although can also invade the inner epidermal and possible the hypodermal cell walls. The cuticle constitutes a protective barrier in water transport, gas exchange and pathogen defense (Dominguez et al., 2011), while the epidermis and the hypodermis, with thick cell walls, provide mechanical stress resistance (Khanal and Knoche, 2014). Along fruit development, changes in skin cell division and enlargement, depositions of new cell-wall

material, and changes resulting from the activity of cell-wall-modifying enzymes, produce changes in skin properties. Differences in cuticle and epidermis growth ratio may cause the rupture of stomata in the epidermis as well as micro cracking. As a healing mechanism, a periderm layer is formed beneath the hypodermis producing cork cells that seal the rupture wound in the case of failed stomata or repair the wounds in the intermediate areas of the epidermal surface. In the first case the cork cells are known as lenticels while in the second as russeting (Khanal et al., 2014).

In certain varieties, the skin has a red coloration produced by the accumulation of anthocyanin pigments in the vacuoles of the cells of the epidermis and hypodermis (Dickinson and while, 1986).

The next tissue is the hypanthium or cortex, formed by parenchymal cells and intercellular spaces. There are two hypotheses about the ontogeny of this tissue: the receptacular, for which the hypanthium derives from the extension of the pedicels and the receptable, and the most supported hypothesis, the appendicular, which posits that the hypanthium derives from the fusion of accessory tissues, including petals, sepals and stamens (Pratt, 2011).

The core (pith) is formed by be the fusion between the floral tube and the ovary, and it is formed by the exocarp (in fusion with the hypanthium), the mesocarp (the flesh of the core) and the endocarp (of ovarian origin). From an internal transverse view of the fruit, in the central region there are five locules derived from five carpels, each carpel containing one to four seeds. The seeds are surrounded by the cartilaginous tissue of the endocarp (**Figure 1B**) (Pratt, 2011).

1.3 Origin and domestication of the apple crop

Archeobotanical studies suggest that trees belonging to the Rosaceae family, such as apple, were domesticated rapidly compared to cereal crops, due to hybridization. During the Miocene, large fruit size and appropriate fruit morphology for consumption favored the seed dispersal by wild animals, which is considered an important fact for understanding the origins of the crop (Spengler, 2017).

Vavilov's and other studies conducted along the XXth century (cited in Cornille et al., 2013), based first on fruit morphology and later in genetic diversity, placed the domestication of the cultivated apple in Kazakhstan. More recently, broad analysis of chloroplast and nuclear genomes of wild and cultivated apples have stablished the origin of the cultivated apple in the wild species *Malus sieversii* (Ldb.) Roem, in the Tian Shan mountains located in Central Asia (border between China and Kazakhstan) (Harris et al.,2002). Domesticated apples were distributed from central Asia to the west (Europe) through the, Silk Road (**Figure 2A**) (Spengler 2017), where coincided and were intercrossed with other wild Malus species such as *Malus baccata* (L.) Borkh. in Siberia, *Malus orientalis* Uglitz. in the Caucasus and *Malus sylvestris* Mill. (European crabapple) in Europe. The contribution of Malus species to the genome of the current cultivated apples (Cornille et al., 2013a; Sun et al., 2020) identify hybridization as a fundamental phenomenon that led to gene introgression throughout domestication (**Figure 2A**). While a relevant secondary contribution of European crabapple to *Malus x domestica* has been proved, molecular markers and nuclear mitochondria sequences have demonstrated a minor contribution of other wild species such as *M. orientalis* (Cornille et al.,2012; Nikiforova et al., 2013). Recently, Sun et al., (2020) found that close to the 23% of the genome of *Malus* x *domestica* cv Gala derives from *M. sieversii* and *M.*

sylvestris, and that hundreds of the genes largely fixed in the pangenome of cultivated apples derive from these two progenitors. In addition, data reveal that interspecific hybridization and gene introgression has also occurred from cultivated to wild. Therefore, we may conclude that hybridization events have shaped the genome of the cultivated as well as the wild apples, probably favored by the self-incompatible mating system o and the cross-compatibility between Malus species. (**Figure 2B**) (Harris et al.,2002).

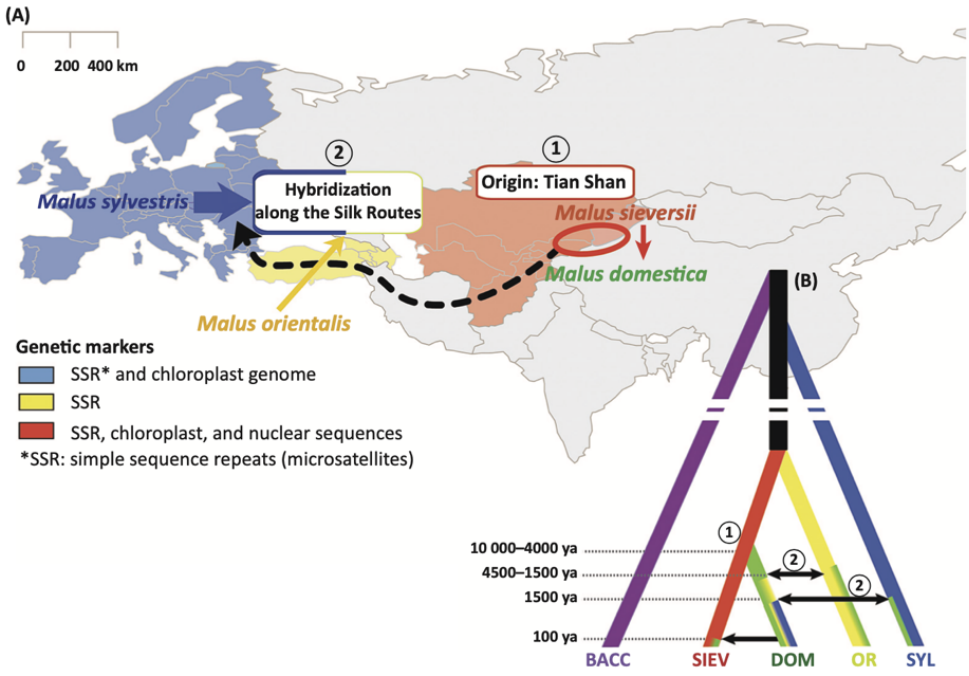


Figure I.2. Evolutionary history of the cultivated apple. **A,** genetic studies with molecular markers (microsatellites) of apple populations in Eurasia have revealed the origin and evolutionary implications of hybridization in the origin of the cultivated apple tree . (1) Origin in the Tian Shan Mountains from *Malus sieversii*, followed by (2) dispersal from Asia to Europe along the Silk Road, hybridization, and gene introgression from Caucasian and European crabapples. **B,** genealogical relationships between wild and cultivated apples. Approximate dates of domestication and hybridization events between wild and cultivated species are detailed in the legend. Abbreviations: BACC, *Malus baccata*; DOM,

M. × domestica; OR, *Malus orientalis*; SIEV, *M. sieversii*; SYL, *Malus sylvestris*; ya, years ago. This image was used from (Cornielle et al 2013a).

1.4 Production and uses

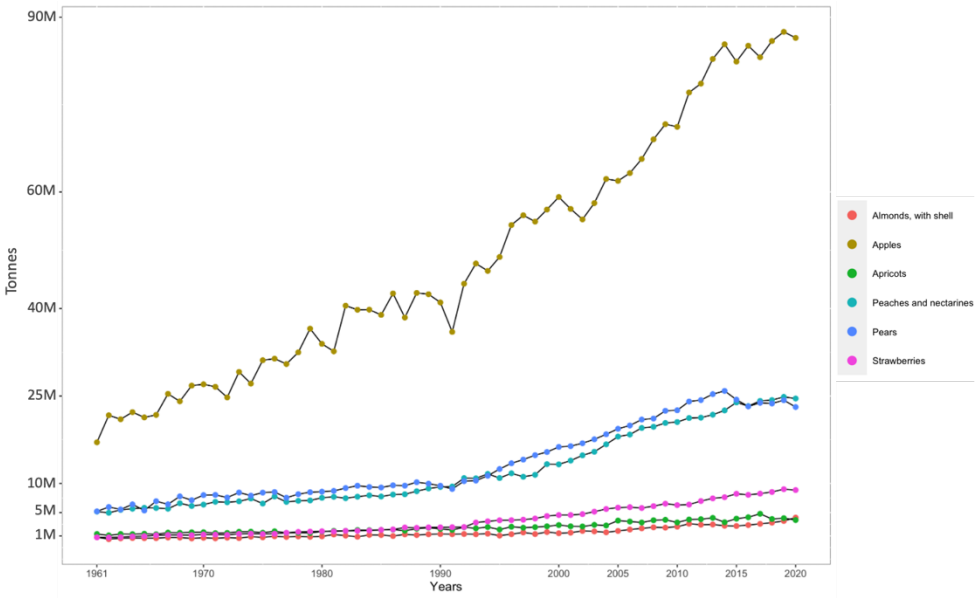
Archeologists found evidence of apple consumption in 35,000-6,500 years AC, proving that crabapples have been part of the diet of animals and humans for thousands of years. As described above, during the development of agriculture and the rise of large cities, apples were traded and cultivated.

In terms of production, apple is among the most cultivated fruit species FAOSTAT (2020), with a much higher production than other fruit crops such as almond, apricots, peaches and nectarines, pears, and strawberries (**Figure 3A**). Although apples are cultivated in all continents, the countries with the highest average production between are China, occupying the first place, followed by the USSR, the USA, France, and Italy. Spain occupies the 17th place (**Figure 3B)** with a harvested area of 29,490 ha and a production of 522,100 tons in 2020, which was reduced compared to the previous year by 116,740 tons. From the point of view of the cultivated regions, Asia and Europe concentrate 81% of the world apple production (**Figure 3C**) (FAOSTAT, 2020).

Approximately 54% of the marketed apple production in the European Union consists of four main cultivars, 'Golden Delicious', 'Gala', 'Red Delicious' and 'Idared' (WAPA, 2019). Although the greatest apple consumption worldwide is as fresh fruit, an important fraction of the apple production is destinated to processed products as ciders, juices, wines, canned sauces, dried or frozen apples, vinegars, jams, and butter among others. The first references to cider, an alcoholic beverage made with fermented apple juice, date to the 55 B.C, in the Roman Empire times, and it is known that cider was already

produced in the northern Spain before the birth of Christ. After the fall of the Roman Empire, the Islamic Moors who ruled a large part of Spain until the end of the fifteenth century A.D, developed new varieties and techniques to produce cider. Wars and conquers favored the distribution of cider from Normandy to England and from there to the English colonies located in North America along the XVIIth century (Watson, 2013). Currently, cider is a very popular drink in Europe, where different types (sparkling sweet, sweet, and dry) are made with appropriate apple varieties (Way and McLellan, 1989). Before World War II most of apple juice was destinated for cider, but in the 1970s its production increased considerable, becoming the second in rank in fruit juice consumption after orange juice (Bump, 1988). Currently there is a large industry dedicated to apple juice production, mainly for clarified juice since consumers prefer totally clear, shining apple juice which requires the removal of materials in suspension and prevention of turbidity after bottling (Kilara and Van Buren, 1989). Such apple industry is mostly based on apple varieties as "Granny Smith", "Fuji", "Gala" and "Braeburn" grown in high-density plots with computerized management systems.

A



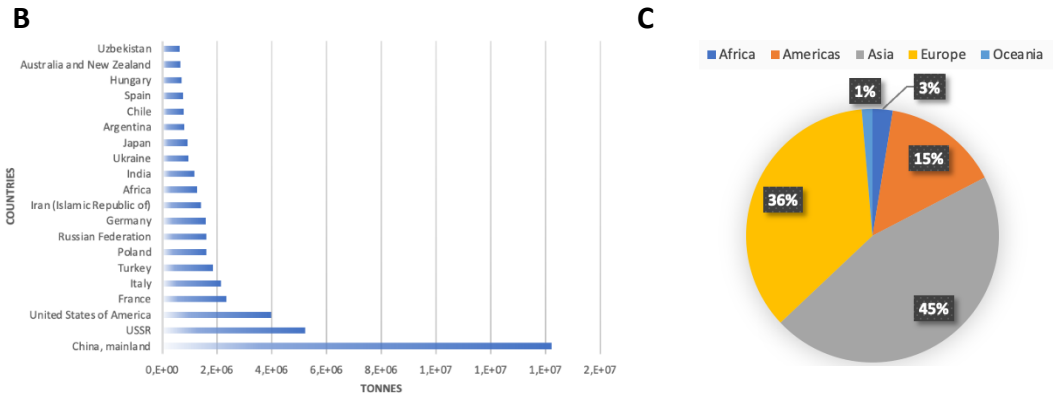


Figure I.3. Production data from FAOSTAT. **A**, comparative graphic of the Production of the major fruits crops in Worldwide. **B**, the 20 top producer countries of apples, the x-axis the unit is million tonnes. **C**, production share of Apples by region. Based on all year reports from FAOSTAT (1961-2020).

SECTION 2: GENETIC AND GENOMIC RESOURCES

Thousands of domestic apple genotypes (both or dessert and cider) have been selected for hundreds of years in Europe, Asia, and North America, and more recently in the southern hemisphere. Together with wild species, these are maintained in national collections as genetic resources for breeding, particularly as sources of resistance to apple scab [*Venturia inaequalis* (Cooke) Winter], powdery mildew [*Podosphaera leucotricha* (Ellis and Everh) Salmon] and fireblight [*Erwinia amylovora* (Burill) Winslow et al.,].

2.1 Genetic diversity of apples

The diversity of Malus species has been explored in both hemispheres. Genetic polymorphisms, intra or inter species, have considerable implications in the evolution and conservation of species. Therefore, their study has relevant application to population genomic studies (Ellegren and Galtier, 2016).

Till the development and use of molecular markers in plants, germplasm was traditionally characterized with morphological and phenological descriptors. The isozymes (described in Stampar and Smole, 1992; Marquard and Chan, 1995) were the first biochemical markers used for this purpose. Soon later, isozymes were substituted by novel techniques based in DNA amplification with universal primers, such as Amplified Fragment Length polymorphism (AFLP) and Random Amplified Polymorphic DNA (RAPD), or even included DNA restriction as the Random Fragment Length Polymorphisms (RFLP) markers. However their applicability for high throughput characterization was limited due to low reproducibility (in the case of RAPD) or high time consuming and cost (Marwal and Gaur, 2020). At present two molecular markers are widely used in breeding programs, microsatellite or SSR (simple sequence repeat) and SNP (single nucleotide polymorphisms). These markers have enabled the development of new breeding strategies, such as marker-assisted selection (MAS), genomic selection, genome-wide association mapping, and high-throughput genotyping, all based on genomic sequences of wild and domesticated species (Velasco et al., 2010; Chagné et al., 2012; Troggio et al., 2012; Bianco et al., 2014, 2016; Gao et al., 2015; Duan et al., 2017; Sun et al., 2020).

In breeding, wild species have been used to introgress certain genes in the cultivated apples providing a diversity of alleles, either to disease resistance, fruit quality, rootstocks, and plant architecture (kumar et al., 2010, Bus et al., 2005; Fazio et al., 2009,2012; Duan et al., 2017). According to the Germplasm Resources Information Network-Global (USDA, 2020), 18 natural hybrids have been identified as source of genetic diversity used in crosses for cultivar improvement. For example, a *Malus* × *robusta* (*M. baccata* × *M. prunifolia*) recombinant line (Markussen et al., 1995), has

been widely use to introduce the Pl1 gene to develop new varieties resistant to powdery mildew. Similarly, *Malus* × *zumi* (*M. mandshurica* × *M. sieboldii*) has been also used as donor of resistance to powdery mildew and of other desired fruit quality traits (Schmidt, 1994).

Several apple conservation programs have been stablished worldwide. For example, the Chinese National Repository of Germplasm Resources in China, center of origin of 17 wild species and 6 native domesticated species, preserved in 2015 more than 1,500 Malus accessions (Gao et al.,2015). Genetic diversity of wild relatives such as *Malus baccata*, *Malus prunifolia*, *Malus* × *robusta*, and *Malus sieversii* of this collection was analyzed by Gao et al (2015) together with 391 accessions of *Malus* × *domestica* from China, Japan, former Soviet Republics, and Western countries using SSRs. The results showed that *Malus* × *domestica* cultivars from former Soviet Republics were more closely related to *Malus sieversii* while Chinese Malus × domestica accessions were closer to those from Western countries than those from Japan (Gao et al., 2015).

Gros et al., (2014) studied the genetic diversity among the most common cultivars used in the United States for their production during the 13th to 20th centuries. They found high genetic diversity, with average expected heterozygosis values (He) higher than 0.7. This data contrasts with the lower He found among the modern varieties, which suggests that the artificial selection conducted in breeding programs could have a slight impact on the genetic diversity of current cultivars. Similar results were found in Potts et al., (2012).

In Europe, a large-scale population analysis was carried out using 2,446 accessions and 16 SSRs. The accessions were provided by eleven countries representing three broad European regions (North + East, West, and South). The analysis revealed a high level of

diversity and heterozygosity in apple germplasm at the European level. This accessions were differentiated in three groups according to their population structure and geographic distribution from the northeast to the South of Europe (Urrestarazu et al., 2016). This analysis also showed that the cultivars most used in Europe were close related and represented low level of genetic variability (Urrestarazu et al., 2016).

2.2 The apple genome

Most Malus species are known to be diploid (2n= 2x= 34), but there are also species that are triploid (2n=3x=51), and tetraploid (2n=4x=68). Evans and Campbell (2002), based on molecular phylogenetic analysis, supported the hypothesis of the autopolyploid origin of the Maloideae from the close taxa Gillenia, which chromosomal number (x) is 9. According to this hypothesis, the aneuploid loss of one pair of chromosomes would have originated the x=17 from the x=18. Later, the studies of Velasco et al., (2010) supported this hypothesis and found that the duplication of gene families involved in fruit development may explain formation of the characteristic pome fruit, developed by proliferation of the receptacle.

Although molecular markers have provided information on sequence similarity between species or cultivars, knowing the structure of the genome is key to understand plant evolution and genome functionalities. The advances in Next Generation Sequencing (NGS) technologies has allowed numerous studies, from phylogeny analysis to knowledge of the genes that can control plant growth and fruit development, from the seedling to maturity, as well as plant adaptation to environment (Peace et al., 2019). The first draft of the apple genome was obtained by Velasco et al., (2010) from the 'Golden Delicious' variety, which was sequenced using a Whole-Genome Shotgun approach, a mixture of Sanger sequencing and Rocher 454 sequencing. The estimated

size of this version of the apple genome was 742.3 Mb with a sequence coverage of 16.9-fold total, 26% of which come from Sanger sequencing and 74% from 454 sequencing. The assembly consisted of 1,629 metacontigs, which represented the 17 chromosomes of which the total contig length (603.9 Mb) covered about 81.3% of the genome. The total number of predicted genes was 57,386. The analysis revealed strong collinearity between chromosomal segments, i.e. several duplications associated with coding regions along each chromosome, indicating recent Genome-Wide Duplication (GWD), which may date back more than 50 million years as a result of the transition from the common ancestor of the Maloidea group with nine chromosomes (Velasco et al., 2010). Li et al., (2016) improved the first assembly of the domesticated apple 'Golden Delicious'. For this, they performed a de novo genomic assembly obtaining 76 Gb (~102 × genome coverage) from Illumina HiSeq data (high-throughput sequencing), and 21.7 Gb (~29 × genome coverage) from PacBio data (long reads sequencing systems). The size of the de novo genome assembly was 632.4 Mb, covering 90% of the estimated genome size (701 Mb), with the size of each N50 contig being a ~6.9 fold improving in length compared to the first version 16.1 Kb. The number of protein-coding genes annotated in this version of the genome was 53,922.

High heterozygosity and duplication of chromosomal fragments in diploid genomes, as in the case of apple (Malus × domestica), together with high density of single nucleotide variants (SNV) and structural variations in the genome, difficult genome assembly (Kajitani et al., 2014). After the Li et al., (2016) version, a group of researchers at the Institut National de Recherche pour l'Agriculture, l'Alimentation et l'Environnement (INRAE) sequenced a line of 'Golden Delicious doubled-haploid #13' known as "GDDH13". The result of this de novo assembly derived from a combination of three

types of technologies: for short (Illumina) and long sequencing reads (PacBio) accompanied with a scaffolding based on optical maps (BioNano) and a high-density genetic linkage map. A consensus map resulted in a 643.2 Mb-sized assembly for the 17 chromosomes with 42,140 annotated protein-coding genes (Daccord et al., 2017).

Currently, two new apple varieties 'Hanfu' and 'Gala' of the species *Malus* × *domestica* have been sequenced. Zhang et al., (2019) obtained the genome of the HFTH1 line of the Chinese variety known as 'Hanfu' (Dongguang × Fuji), offspring of two parents of great interest for their desired fruit quality. As a result of a combination of single-molecular real time (SMRT) sequencing (PacBio), chromosome conformation capture (Hi-C) sequencing, and optical mapping, they assembled a high-quality genome of 658.9 Mb with an N50 contig of 6.99 Mb, 44,677 protein-coding genes annotated, 18,047 deletions, 12,101 uncertainties and 14 long inversions (due to high TEs activity). In addition, they compared the TE with the ones in the GDDH13 genome, identifying the 59.8% in HFTH1 (Zhang et al., 2019). Recently the diploid variety 'Gala' (*Malus* × *domestica*) has been sequenced together with other wild species such as *M. sieversii*, *M. sylvestris*. In the case of the domesticated variety, they obtained a 623-780x coverage with Illumina and 10x Genomics sequences and an additional coverage of 37-81x with PacBio HiFi sequencing. The estimated size of the diploid genome with phased scaffolds was 1.31-1.32 Gb while for the haploid consensus was 652-668 Mb, with the N50 scaffold of 16.8-35.7 Mb. For this genome the number of protein-coding genes annotated in the haploid consensus was 45,199-45,352 (Sun et al., 2020).

The genomes of wild species, such as *M. baccata* (Chen et al., 2019), *M. sieversii* and *M. silvestris* (Sun et al., 2020) *M. prunifolia* (Li et al., 2022), have also been obtained and are publicly available on databases such as the Genome Database for Rosaceae (GDR),

as described by Jung et al., (2019). In addition, apart from the four species mentioned above, Duan et al., (2017) sequenced 17 wild species. While two of the four species have contributed to apple tree domestication, wild species may also represent a source of allelic diversity to tolerate abiotic or biotic stress which can be used in a breeding program.

The first apple genome sequence gave a high knowledge about gene discovery, polymorphic variants, and was extremely useful for the development of marker-arrays. Subsequent versions of the apple genome have been released, with GDDH13 v1.1 being the reference genome for its high quality, which has provided better biological and physiological understanding of interesting traits. At the same time, the high genetic resolution has been the basis for the development of high-throughput platforms for breeding programs. It also served to elucidate, with more confidence, the origins of domestication and the influence of wild relatives on the domestication event (Peace et al., 2019).

SECTION 3: MARKER-TRAIT ASSOCIATION AND MARKER-ASSISTED BREEDING IN APPLE FOR FRUIT QUALITY

3.1 Genetics maps & Marker-assisted breeding

The relative position within the chromosomes of molecular markers and sequence data can be calculated through genetic and physical maps. The genetic or linkage maps, show the location of the genes and the distance between them expressed in centiMorgan (cM), while in the physical maps, the distance between genes, variants and other DNA sequences of interest, is expressed in base pairs (bp) (O'Rourke, 2014).

Several platforms and projects have been developed (Teh et al., 2q021) to aid apple breeding. In Europe, international collaborative projects such as HiDRAS and

FruitBreedomics, aimed at developing new tools for breeding and at bridging the gap between research and breeding. The main outputs of these projects were: 1) for the HiDRAS project, the construction of 13 linkage maps as well as an apple integrated map made of six F1 populations totaling 720 individuals and including 1750 markers along the 17 chromosomes (Gianfranceschi and Soglio, 2004; Velasco et al., 2010), and 2) for the FruiBreedomics project, the development of trait-associated molecular markers, the development of phenotyping methods, as well as a broad analysis of the European apple diversity (Laurens, 2010). In the USA, the RosBREED and RosBREED2 projects successfully developped a DNA-based plant breeding platform for Rosaceae species that includes phenotyping protocols, tools for breeders, and a DNA testing portal (Lezzoni et al., 2017). In addition, all these efforts, served to develop pre-breeding materials. As a conclusion, these international and collaborative efforts provided important resources to support breeding, through tools and materials that can increase efficiency, accuracy, and reduction of breeding time.

To the present, research efforts have developed a considerable number of molecular markers and linkage maps related to pathogen resistance, abiotic stress, fruit quality, among others. In terms of fruit quality, there are different interesting traits related to the characteristics desired by breeding programs and, consequently, by the final consumer, such as acidity, sweetness, volatile compounds, phenols, vitamins, multiple textural sensory traits (i.e., chewiness, crunchiness, juiciness, mealiness, flesh melting index and skin thickness), skin color, skin roughness, fruit size and shape, among others (Han and Korban, 2021). La Belle (1981) described the characteristics of ripe fruit for high-quality processing: maturity, damage, fruit shape, decomposition, skin color, pulp

color, firmness, soluble solids, total acidity, pH, organic flavor compounds, tannins, oxidation and juiciness among others.

Numerous QTLs (quantitative trait loci) for fruit quality have been detected in segregating populations and accessions. **Table 1** describes most of the fruit quality traits and parameters that have been currently mapped (Han and Korban, 2021). They can be classified under two categories: **external quality**, related to parameters of size, fruit shape, texture, skin color, firmness, among others, and **internal quality** includes organic compounds, such as sugars, esters, alcohols, acidity-related compounds, terpenes associated for human health benefits, and compounds for fruit storage preservation. Most of the QTLs mapped for fruit quality are associated to the internal quality, such as soluble solids content (SSC) Guan et al., (2015), sugars (Ma et al., 2016, Larsen et al., 2019), acidity (Verma et al., 2019, Rymenants et al., 2020), esters, phenyls, sesquiterpenes (Costas et al., 2013, Cappellin et al., 2015a, Kumar et al., 2015c), triterpenes (Christeller et al., 2019), ethylene production (Costa qet al., 2005, Cappellin et al., 2015b), and health-promoting compounds (McClure et al., 2019) among others. For external fruit quality, the most studied traits are the skin color (Chagné et al., 2016; Migicovsky et al., 2016; Amyotte et al., 2017; McClure et al., 2018; Jung et al., 2020, 2022), texture (Ben Sadok et al., 2015; Amyotte et al., 2017) and size and shape (Liebhard et al., 2003; Kenis et al., 2008; Chang et al., 2014; Potts et al., 2014; Cao et al., 2015).

Table I.1. QTLs reported for apple fruit quality. Modified from Teh et al., (2021)

Trait	Categories	Chromosome	Population	Reference
Skin color	External Appearance	9	85 cultivars	Amyotte et al., (2017)
Soluble solids content (SSC)	Internal quality	8		
Fresh green apple	Internal quality	9, 12		
Crispness	Internal quality	5, 13		
Juiciness	Internal quality	13		
Mealiness	Internal quality	5, 10		
Skin thickness	Internal quality	10		
Fibrousness	Internal quality	7, 10	X3259 × X3263	Ben Sadok et al., (2015)
Firmness	Internal quality	10, 11		
Crunchiness	Internal quality	10		
Graininess	Internal quality	1, 7, 10		
Mealiness	Internal quality	1, 2, 4, 7, 12		
Meltiness	Internal quality	5, 6, 7, 8		
Juiciness	Internal quality	1, 11		
Fruit shape index	External Appearance	11	Jonathan × Golden Delicious	Cao et al., (2015)
Ethylene	Organic compound	3, 13, 15	Golden Delicious × Scarlet	Cappellin et al., (2015a)
Estragole / Propanal Butanal	Organic compound	17 / 16		
1-butanol	Organic compound	3		
Alcohols and esters / Farnesene	Organic compound	8 / 5		
Acetate esters	Organic compound	2, 15	Fuji × Delearly 124 accessions	Cappellin et al., (2015b)
Esters	Organic compound	2, 4, 5, 15		
Ethylene	Organic compound	2, 14		
Furanes	Organic compound	4, 5, 13		
Phenyls	Organic compound	3		
Sesquiterpenes	Organic compound	5		
Skin overcolor	External Appearance	9	4 full-sib families	Chagné et al., (2016)
Fruit Diameter	External Appearance	2, 5, 8, 13	Jonathan × Golden Delicious	Chang et al., (2014)
Fruit Height	External Appearance	4, 8, 11, 15, 17		
Fruit Shape Index	External Appearance	4, 13, 15		
Fruit Size	External Appearance	2, 8, 11, 12, 14, 15		
Triterpenes	Organic compound	3, 5, 9, 17	Royal Gala × Granny Smith	Christeller et al., (2019)
Ethylene production—Md-ACO1	Organic compound	10	Prima × Fiesta	Costa et al., (2005)
Ethylene production—Md-ACS1	Organic compound	15	Fuji × Mondial Gala Fuji × Braeburn	
Expansin (softening)—Md-Exp7	Organic compound	1	Prima × Fiesta 31 cultivars	Costa et al., (2008)
Esters	Organic compound	2	Fiesta × Discovery	Costa et al., (2013)

Ethylene	Organic compound	15		
Alcohols Esters	Organic compound	2, 3 2, 3, 9	Discovery × Prima	Dunemann et al., (2009)
Volatile organic compound profiles	Organic compound	2, 10	162 accessions	Farneti et al., (2017)
Fructose	Organic compound	1, 3, 15	15 families and 41 accessions	Guan et al., (2015)
Glucose	Organic compound	1, 2, 3, 15, 16		
Sucrose	Organic compound	1, 3, 4, 9, 12		
Sorbitol	Organic compound	1, 3, 5, 9, 11, 13, 15		
Soluble solids content	Internal quality	2, 3, 12, 13, 15		
Soft scald	External Appearance	2, 16	4 full-sib families	Howard et al., (2018)
Fruit diameter	External Appearance	2, 5, 10, 17	Telamon × Braeburn	Kenis et al., (2008)
Fruit height	External Appearance	2, 6, 17		
Stiffness	External Appearance	16		
Mean firmness of red/sun side	External Appearance	10		
Mean firmness of green/shaded side	External Appearance	10		
BrixR:Mean SSC of red/sun side	Internal quality	10		
BrixG:Mean SSC of green/shaded side	Internal quality	10		
Acidity	Internal quality	16		
Quercetin conjugates	Organic compound	1, 13	Prima × Fiesta	Khan et al., (2012)
Skin phenolics	Organic compound	16		
Flesh phenolics	Organic compound	16		
Compression	External Appearance	1, 16	Prima × Fiesta	King et al., (2001)
Maximum force	External Appearance	16		
Wedge fracture	External Appearance	1, 6, 8, 15		
Specific gravity	External Appearance	6		
Compression stiffness modulus	External Appearance	6		
Fruit firmness	External Appearance	10	7 full-sib families	Kumar et al., (2013)
Weighted cortical intensity Internal browning	External Appearance	9		
Titratable acidity	Internal quality	8		
Fruit splitting	External Appearance	16		
Alcohols	Organic compound	2, 15	230 accessions	Kumar et al., (2015c)
Terpenes	Organic compound	2, 12		
Acetate esters	Organic compound	4, 8		
Ethyl esters	Organic compound	17		
Other esters	Organic compound	1, 15, 17		
Acetate esters	Organic compound	2	145 Danish heritage apple cultivars	Larsen et al., (2019)
Sucrose content	Organic compound	1		

% sucrose	Organic compound	1		
% fructose	Organic compound	1		
Flesh firmness	External Appearance	3, 6, 12	Fiesta × Discovery	Liebhard et al., (2003)
Fruit weight	External Appearance	6, 16		
Size	External Appearance	8, 17		
Acidity	Internal quality	8, 16		
Polygalacturonase texture—Md-PG	Organic compound	10	Fuji × Delearly Fuji × Cripps Pink	Longhi et al., (2012)
Polygalacturonase texture—Md-PG	Organic compound	10	77 cultivars	Longhi et al., (2013)
			Fuji × Delearly	
Fructose	Organic compound	3	Jiguan × Wangshanhong	Ma et al., (2016)
Glucose	Organic compound	3, 4		
Sucrose	Organic compound	3		
Sorbitol	Organic compound	3		
Malic acid	Organic compound	8, 16		
Fruit firmness	External Appearance	1, 10	Prima × Fiesta	Maliepaard et al., (2001)
Soft scald	External Appearance	2, 3	11 W-12–11 × SPA440	McClure et al., (2016)
			Ambrosia × Honeycrisp	
Fruit skin color	External Appearance	9	172 accessions	McClure et al., (2018)
Change in firmness	External Appearance	10	172 accessions	
Catechin Epicatechin Procyanidin B1 Procyanidin B2 Procyanidin C1	Organic compound	16	136 accessions (in 2014)	McClure et al., (2019)
Quercitrin	Organic compound	1	85 accessions (in 2016)	
Chlorogenic acid	Organic compound	5, 15		
4-O-caffeoylquinic acid	Organic compound	3, 14		
Cyanidin-3-galactoside	Organic compound	9		
Fruit flesh firmness, Fruit overcolor	External Appearance	3, 9	689 accessions (data mining from USDA-GRIN)	Migicovsky et al., (2016)
Overcolor intensity	External Appearance	9		
Circumference	External Appearance	3, 5	Co-op 17 × Co-op 16	Potts et al., (2014)
Diameter	External Appearance	3		
Length	External Appearance	3, 5		
Weight	External Appearance	3, 5		
2-methylbutyl acetate	Organic compound	2	Royal Gala × Granny Smith	Rowan et al., (2009)
Sensorial acidity: Ma, Ma3, Ma4, Ma5	Organic compound	16, 8, 6, 1	3 full-sib families	Rymenants et al., (2020)
Sensorial sweetness	Internal quality	8, 15, 16		
a-farnesene	Organic compound	5, 10, 12, 15	Royal Gala × Granny Smith	Souleyre et al., (2019)
Malic acid	Organic compound	8		Sun et al., (2015)

Citric acid	Organic compound	8, 15	Jonathan × Golden Delicious	
Acetic acid	Organic compound	7		
Total acid	Organic compound	8		
Fructose	Organic compound	1		
Sucrose	Organic compound	1		
Fruit weight	External Appearance	3, 5		
Fruit firmness	Internal quality	11		
Titratable acidity:Malic acid—Ma, Ma3	Organic compound	16, 8	16 full-sib families	Verma et al., (2019)
Lipoxygenases	Organic compound	2, 4, 5, 6, 7, 9, 11, 12, 13, 16	Discovery × Prima	Vogt et al., (2013)
Esters	Organic compound	2, 9, 12		
Hexanals	Organic compound	7, 12		
Malic acid—Ma locus: Titratable acidity	Organic compound	16	Royal Gala × PI 613,971	Xu et al., (2012)
pH	Internal quality	16	Royal Gala × PI 613,988	

The research of this Thesis is focused on fruit morphology, including shape and size. Fruit size and shape are polygenic traits with an important environmental component (Daccord et al., 2017). At the present, these traits have been considered in few studies, and using segregating populations, reducing, thus, the genetic variation to that present in the parental lines.

For example, Kenis et al., (2008) studied the inheritance of fruit size in a population of 'Telamon × Braeburn', detecting two QTLs in LG10 and LG15 for fruit diameter and fruit length, respectively that explained the 22-33% of the phenotypic variation. Later, Devoghalaere et al., (2012) hypothesized about a putative role of the Auxin Response Factor gene (*MdARF106*), detected in the regions of one LG15 QTL, in cell expansion, and, therefore, in fruit size control. Also, Yao et al., (2015) detected four QTLs one of which (in LG11) co-locating with a microRNA (miRNA172) fixed in cultivated apples. which over-expressions of this miRNA correlated with lower fruit size; the effect was validated in apple transgenic lines.

Chang et al., (2014) also detected several QTLs for fruit shape index (FSI), one of those QTLs in LG11 contributed to a phenotypic variance of 13.7% in a segregating population. Later, Cao et al., (2015) analyzed the QTL-LG11 region of the same population studied, identifying a candidate gene (LysM domain receptor-like protein kinase) related to a non-synonymous SNP in the population.

An OpenArray v1.0 assay was developed by the International RosBREED Consortium for apple (IRSCOA v1.0). It includes 128 SNP-type molecular markers associated to fruit quality, pest, and disease resistance traits. Thirty-three markers from IRSCOA v1.0 have been validated for use in marker-assisted selection (MAS) using commercial materials, elite selections and segregating populations forming part of the Plant and Food Research, New Zealand breeding program. These validated markers are for scab, fireblight, powdery mildew resistance allele and for fruit quality such as, firmness, skin color, flavor intensity and acidity (Change et al., 2019).

3.2 Genome Wide Association Studies (GWAS)

The objective of genome-wide association studies (GWAS), also known as genome-wide association mapping, is to detect the association between the genotypic frequency and the trait observed in a population of unrelated individuals. For that, the methodology requires two matrices, one with the phenotypes and one with the genotypes. The first, represents the population according to the trait under study. For example Larsen et al., (2019) studied the variability of volatile compounds within a germplasm collection of Danish heritage cultivars. The genotypes matrix consists of the genotypic data of the population obtained by either SNPs arrays, genotyping by sequencing (GBS), whole genome sequencing (WGS) or combined by imputation (Tam et al., 2019). Following acquisition of both matrices, association statistics are applied by means of mathematical

models such as MLM (Zhang et al., 2010), MLMM (Segura et al., 2012), FarmCPU (Liu et al., 2016) and BLINK (Huang et al., 2019) were some parameters as kinship or population structure indexes can be included. The result of this analysis will provide the significant SNPs of the association, revealing the region of the QTN (quantitative trait nucleotide), considering the linkage disequilibrium of the region detected in the set of individuals analyzed.

The design and the marker density required for the genotypic array largely depends on the diversity of the species and, in major extent, to the linkage disequilibrium (LD), i.e., the distance between markers at which are inherited independently in the population of study. Apple is a highly variable species; therefore, the design of a SNP array should take into account the variability in the regions flanking the SNP and select alleles with relatively high frequency (since low frequency alleles will be removed from the analysis). Also, LD in apple decays fast (two markers 2-2.5 kb apart tend to be inherited independently, r^2<0.2) (Urrestarazu et al., 2017; Jung et al., 2020).

Four apple, SNPs arrays such as the 8K (Chagné et al., 2012), 20K (Bianco et al., 2014), 480K (Bianco et al., 2016), and 50K (Rymenants et al., 2020) have been developed and are available for their use. However, considering the apple LD, only those with higher density are suitable for GWAS analysis.

The European FruitBreedomics Consortium selected a collection representing the diversity of apple cultivars and current breeding materials in Europe, called the apple REFPOP, which has 534 genotypes (consisting of 269 accessions and 265 progenies from 27 parent combinations) planted in six European countries (Belgium, Spain, France, Italy, Poland and Switzerland). The collection was densely genotyped, with either the 480K SNP array (in the case of the accessions) or with the 20K array (in the case of the

progenies). The analysis and data imputation required to combine the data from the two arrays produced a high-genetic matrix of 303,239 SNPs.

The high-density SNP data was used by Jung et al., (2020) to study diversity as well as population structure of the REFPOP. A neighbor-joining tree showed that the non-European accessions were grouped in the upper part of the clade, and below them were the remaining accessions, being predominantly those from Western and Central Europe (WCE) (Fig. 4A). Also, the principal component analysis showed that the PC1 + PC2 explained the 7.6 % of the total variance, indicating that the progeny did not formed separate clusters, while the accessions were dispersed on both axes (Fig. 4B). Finally, by means of the ADMIXTURE analysis, it was concluded that the population structure of the REFPOP apple was weak and with a high level of admixture, as shown in (Fig. 4C) the genotypes defined according to the geographic region of origin are highly mixed (Jung et al., 2020).

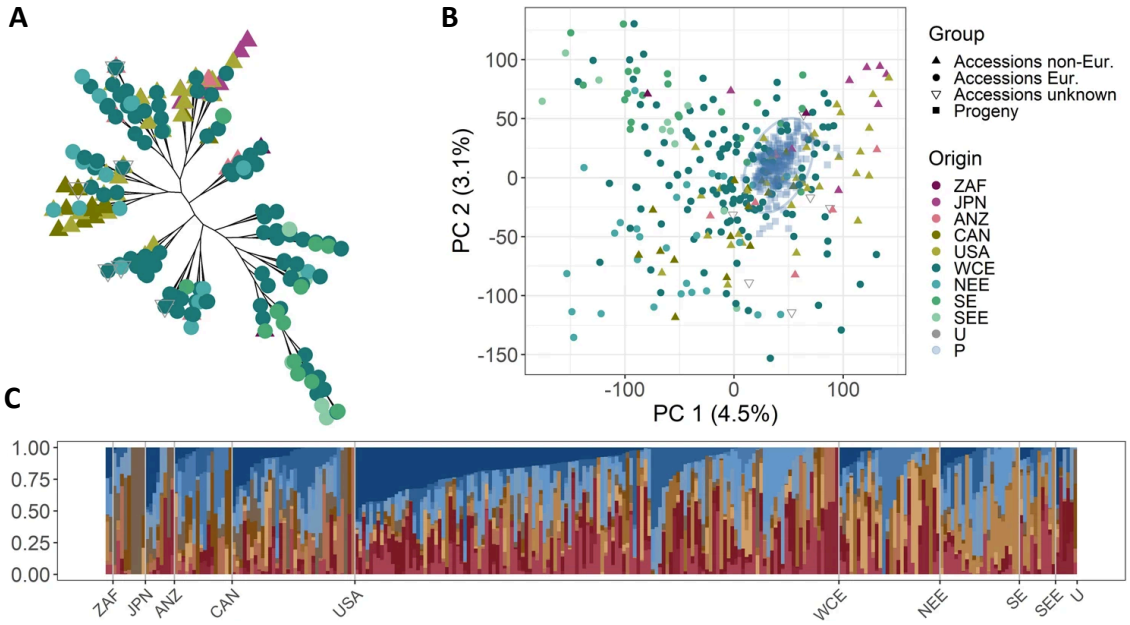


Figure I.4. Structure of the apple reference population. **A** Unrooted neighbor-joining tree of the accession group, colors correspond to the legend in "B". **B** Principal component analysis of the accession group with progeny group as supplementary

individuals encircled with a normal confidence ellipse. The total variance explained by two components. **C** ADMIXTURE bar plot of the accession group. Labels in plots "A" to "C" refer to the geographic origin of genotypes: ZAF (South Africa), JPN (Japan), ANZ (Australia and New Zealand), CAN (Canada), USA (United States of America), WCE (Western and Central Europe), NEE (Northern and Eastern Europe), SE (Southern Europe), SEE (Southeastern Europe), U (accessions of unknown geographic origin), and P representing the progeny group in plot "B". The plots and analysis were used from Jung et al., (2020).

In terms of linkage disequilibrium (LD), a fast decay was observed in concordance with data obtained in other studies (Urrestarazu et al, 2017) (Fig. 5A), with r^2<0.2 at 2.52 kb distance (Fig. 5B) (Jung et al., 2020).

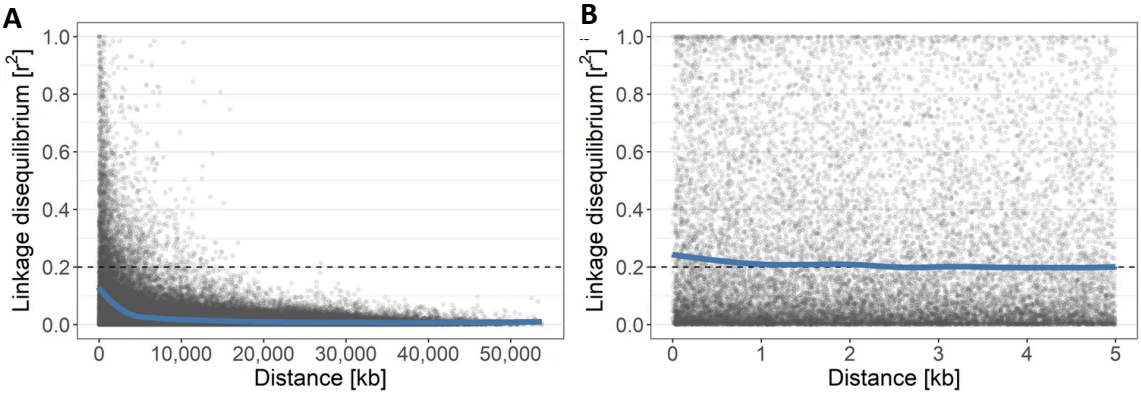


Figure I.5. Linkage disequilibrium decay in the apple reference population. Linkage disequilibrium with a loess smoother for **A** distances between SNPs across the span of chromosomes, and **B** for SNPs within a 5 kb distance The plots and analysis were used from Jung et al., (2020).

The multi-environment design of the REFPOP allows the study of the phenotype-genotype effect, as well as the environmental effect on the trait. GWAS as well as genome prediction models will help to identify future interesting parentals as well as to increase the gain in selection processes (Jung et al., 2020).

During the course of this Thesis, we (the research group of Dr. Aranzana) participated in the characterization of phenology, production and fruit quality of the REFPOP, which was used to conduct GWAS and genomic prediction on 30 traits. The study revealed that

the genotype-environment interaction (G × E), when considering the six REFPOP hosting locations, represented up to 24% of the phenotypic variability. Between the traits studied, QTNs for fruit size parameters were detected (Jung et al., 2022).

However, in this Thesis, an exhaustive analysis of morphological parameters of the fruit has been carried out, obtaining numerous associations (QTNs), which has been contrasted with already published (see **Chapter 1** and **Chapter 2**).

SECTION 4: FRUIT DEVELOPMENT

4.1 Fertilization, seed formation and fruit development

During the reproductive process of plants, mechanisms such as the self-compatibility, semi-incompatibility, and self-incompatibility determine the fertilization. The pollination process starts when a pollen grain contacts the floral stigmatic surface, germinates due to the extracellular secretions, and grows into the pollen tube to, finally, fertilize the egg. In self-incompatible (SI) systems, as in most rosaceous, complex processes mediated by proteins and specific molecules recognize the pollen and allow or prevent its growth and further germination of the ovule, preventing inbreeding. This recognition process occurs in the gametophytic and sporophytic system, in which the inhibition of the pollen tube growth is controlled by the multi-allelic S locus through the involvement of ribonucleases and *F-BOX* proteins. Therefore, when the S haplotype of the pollen and the pistil are different, i.e. are compatible, the recognition complex accepts the growth of the pollen tube through the pistil so the germinated pollen fertilizes the ovule. When one of the S haplotypes is the same in the pollen and pistil, the stigma activates the ribonuclease (S-RNase) degrading the pollen tube; the system is called semi-compatible. But, when both S-haplotype from pollen and pistil are the

same, there are not recognition by the pistil and the pollen tube is degraded; the system is called self-incompatible (Matsumoto et al., 2014).

In some self-incompatible varieties, like 'Cox's Orange Pippin', some self-pollen tubes are semi-compatible since can penetrate the ovary but, in natural conditions, their slow growth prevents egg fertilization. However, factors as favorable temperature conditions (below 25ºC) can favor a faster growth and allow such fertilization. This is known as pseudo-compatibility (Williams and Maier, 1977). However, fruit set under these conditions is highly reduced, for only 1% to 11% of the pollen tubes reach the eggs (Williams and Maier, 1977; De Witte et al., 1996).

The use of pollinators as pollen vectors of compatible cultivars during the bloom is required for fruit setting. Currently, the cross-pollination is implemented in breeding programs to increases productivity. In addition to known the specific S-haplotypes using molecular techniques, allows selecting cultivars that are compatible with each other (Schneider et al., 2005; Larsen et al., 2016).

After pollination and fertilization, there is a transition from flower organs to fruit, known as fruit set, leading to fruit and seed formation (Eccher et al., 2014). Janssen et al., (2008) describes the development of apple fruit within a period of 150 days from full bloom to ripe fruit, performing 8 sampling time points and analyzing the physiological events during development. The first point 0 days after anthesis (DAA) the flower is completely open, the stigma interacts with the pollen for subsequent fertilization of the ovules, triggering cell division until 35 DAA, but between 14 and 25 DAA coincides with the start of the cell expansion phase, in which the rate of cell expansion increases and at 35 DAA begins the accumulation of starch. At 60 DAA the highest rate of cell expansion is recorded and at 87 DAA it has decreased and continues until full maturity, also at this

point the starch levels decrease and at 132 DAA sugars in the fruit increase and the color of the skin is changing. At 146 DAA, ripening is full and the fruit takes on a strong color, it is estimated that all the starch has been converted to sugars and the softening of the flesh is detectable (Janssen et al., 2008) (Fig. 6).

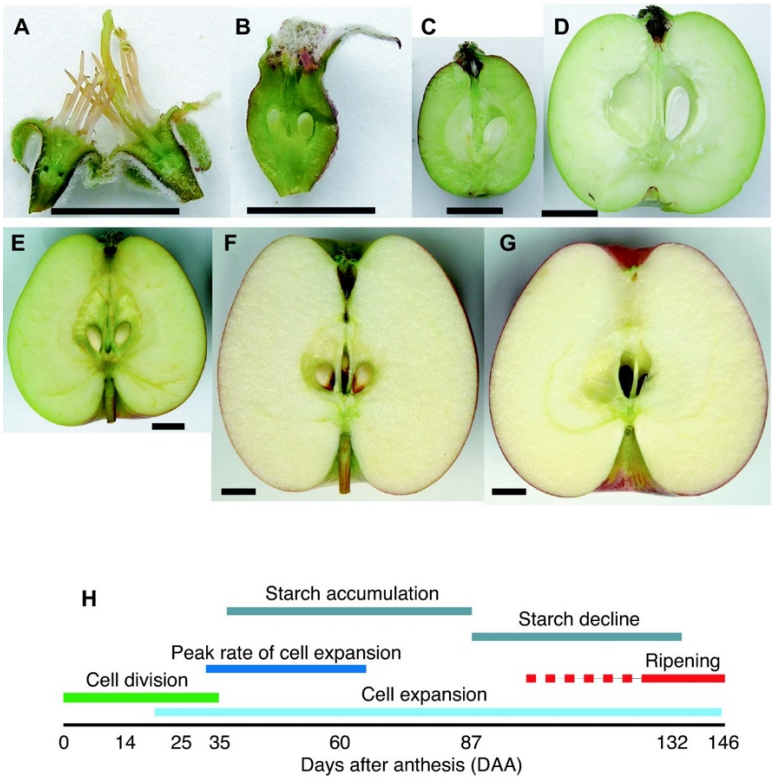


Figure I.6. Apple fruit at stages of development. **A**, 0 DAA, **B**, 14 DAA, **C**, 35 DAA, **D**, 60 DAA, **E**, 87 DAA, **F**, 132 DAA, **G**, 146 DAA. **H**, Scheme of physiological events during fruit development during 0 to 146 DAA. Bar = 1 cm. The images were used from Janssen et al., (2008).

In chapter 3, the study of fruit development in 9 genotypes is detailed, representing three typologies of fruit shape from post anthesis to harvest, obtaining results of cell count and area, RNAseq and DNA-Methylated data.

4.2 Physiology of the apple fruit development.

The physiological factors regulating fruit development are still not well known in apple (Malladi and Johnson, 2011). Some of the fruit tissues are known derive from floral

organs. In flowers of dicot plants, the “ABC model” describes the interaction of homeotic genes to finally stablish the identity of the basic organs (Ma and dePamphilis (2000). Most of these genes are transcription factors, such as the *MADs-box* identified in apple (Yao et al., 1999). High expression of the *APETALA2* (*AP2*) gene in apple (class A) has been found in sepals (Yao et al., 1999), and also in the cortex/flower tube during early fruit development (Kotoda et al., 2000). In the case of the *AP2* gene, which defines the identity of the sepal, it is regulated by microRNA 172 (miR172). This miR172 was identified as a regulator of fruit growth and final size in transgenic plants of the 'Royal Gala' variety (Yao et al., 2015).

In a parthenocarpic apple mutant, Yao et al., 2001 found an increase of the expression of the *PISTILLATA* (*PI*) gene in petals, while not in the other floral organs. Later, Yao et al., 2018 obtained transgenic apple plants with ectopic expression of the PI. The resulting plants produced altered sepals and fruits with reduced growth and modified shape.

In addition, this data suggest that the petals do not contribute to the development of the cortex/floral tube and that the basal regions of the floral organs, especially the sepal, contribute to the subsequent growth of the fleshy part of the fruit. These data support the appendicular theory of fruit development in apple (Malladi, 2020).

During the first stages of fruitset, fruit growth is triggered by a cell production, which at some points stops in favor to cell elongation. Transcriptomic studies have identified a correlation between the amount of 14 *cyclin-dependent kinases* (CDKS) transcripts and the relative cell production rates (*RCPR*) in fruits along development. Since CDKs regulate phase progression during cell mitotic division, these results suggest that the

switch from cell division to cell enlargement may be due to a limitation of available CDKs (Malladi and Johnson, 2011).

In tomato, auxin (AUX) and gibberellin (GA) hormones regulate the fruit set. GA induction promotes cell expansion and AUX promotes cell division during early fruit development (Serrati et al., 2007). In apple, unlike GA, auxin application does not always induce parthenocarpy fruit development (Watenabe et al., 2008). In addition, the application of GA is always accompanied by a change in fruit shape. In experiments with GA application, there was an effect on growth along the polar diameter and it was associated with an increase in the number and size of cells at the distal end of the apple cortex (Nakagama et al., 1968). Natural AUX level in the apple cortex increases during the cell division phase and reaches its maximum in the early phase of cell expansion (Devoghalaere et al., 2012). Similar GA induction experiments in pear have shown an increase in auxin transport and a decrease in ABA biosynthesis (Liu et al., 2018).

In addition to AUX and GA, abscisic acid (ABA) and ethylene are up-regulated at the same time as GA signaling in abscission-induced fruit (Malladi, 2020). After the abscission of some fruitlets on the tree, the cell expansion phase begins, characterized by a strong water absorption through the regulation of the *MdPIP1* gene, a protein located in the plasma membrane (Hu et al., 2003), as well as the transport of sugars and other molecules into the vacuoles. Consequently, fruit volume and starch accumulation increase until maturation Janssen et al., (2008).

As it characterizes the start of ripening in apple, the function of ethylene is crucial for maturation and ripening. Ethylene biosynthesis is based on two systems: in system 1, the genes *MdACS3* and *MdACO3* catalyze biosynthesis, activating it in all vegetative tissues and in unripe fruit (Wiersma et al., 2007; Yang et al., 2013), while in system 2,

the genes *MdACS1* and *MdACO1* catalyze biosynthesis and are responsible for ethylene production and self-stimulated ethylene biosynthesis in the fruit (Gapper et al., 2013; Tan et al., 2013). Although ethylene biosynthesis plays an important role in ripening, the level of ethylene also affects the final fruit quality, for example, acidity and starch degradation during the transition from maturation to ripening are sensitive to this hormone (Johnston et al., 2009). In the final stage of development, numerous genes are expressed, from the degradation of starch in sugars up to the synthesis of aromatic and volatile compounds are regulated by various hormones, finalizing the development of the fruit Janssen et al., (2008).

Fruits size correlates with cell proliferation rather than with cell size (Harada et al., 2005). Daccord et al., (2017) performed an extensive analysis of genomic, transcriptional, and DNA-methylated during early fruit development, studying the 'Golden Delicious' variety GDHH13 and an isogenic line, with smaller fruit, GDHH18. They found a reduction in the number of parenchyma cells in the cortex or hypanthium in the GDHH18 fruits. They also found 22 differentially methylated genes when comparing the two lines, including several transcription factors and a gene associated with ethylene biosynthesis. Many of these genes are homologous to fruit growth regulation genes, such as *APETALA1*, *AGAMOUS-LIKE 8*, *SEPALA1, ETHYLENE-INSENSTIVE3-LIKE3*, among others.

OBJECTIVES

The main objective of this thesis is to study the natural variability of apple fruit size and shape in a wholistic way, combining morphometric, histologic, SNPs, whole genome DNA and RNA data to ultimately increase the knowledge on the genetics and genomics behind apple fruit shape and size determination, and generate new tools for cultivar characterization and breeding.

To achieve this ambitious objective, we addressed three sub-objectives:

1. Exhaustive morphometric description of apple fruit size and shape attributes and use of machine-learning classification methods to determine their weight in class assignation.
2. Genome Wide Association Studies for size and shape measures in apple fruit.
3. Genetic study of fruit shape along apple development from a morphologic, histologic, and differential gene expression perspective, in three fruit shape typologies.

Each sub-objective corresponds to one chapter of this thesis document.

CHAPTER 1

Exhaustive morphometric description of apple fruit size and shape attributes and use of machine-learning classification methods to determine their weight in class assignation.

Dujak, Christian[1], Jurado, Federico[1] and Aranzana, Maria José[1,2]

[1] Centre for Research in Agricultural Genomics (CRAG) CSIC-IRTA-UAB-UB, Campus UAB, 08193, Bellaterra, Barcelona, Spain.

[2] IRTA (Institut de Recerca i Tecnologia Agroalimentàries), 08140, Caldes de Montbui, Barcelona, Spain.

ABSTRACT

So far variety testing is mostly based on human visual inspection. An efficient evaluation method of apple fruit shape would have important applications for breeding and for cultivar characterization. In addition, if this efficient method provides measures of the multiple attributes embraced behind the global term “shape”, it will have an extremely useful application for genomic studies. In this work, we evaluated the use of a shape analyzer software developed to study tomato fruits (Tomato Analyzer) for the analysis of apple size and shape attributes. We analyzed with the software close to 13,000 2D images of apple sections of 364 genotypes collected in three harvest seasons. In addition, we assigned the images into three classes by visual inspection (spheroid oblate or flat, spheroid round and spheroid oblong) as well as into the classes determined in the ECPGR guides. In most of the images the software detected the contour with vague precision, especially in the stalk and calix regions. After visual inspection and manual correction of the contours, we obtained 15 measurements of shape and size attributes. The coefficient of variation of these traits across years ranged from 2.3 to 23.3 %. Numeric values as well as class descriptors were included in a Random Forest model to identify the most important variables determining fruit shape. The fruit shape index (FSII) outstood in importance, followed by the fruit shape triangle (FST), the distal angle Macro (DAMa), the eccentricity (ECC), and the proximal angle macro (PAMa). Shall an efficient image detection method exist, the inclusion of these parameters in fruit description guides would provide more precis descriptions of apple cultivars. In addition, it will be worth to explore a possible genetic control of these traits through genomic studies.

Keywords: fruit shape, apple cultivars, random forest

INTRODUCTION

Apples traded in the international fresh market require to meet fruit quality standards as defined by the OECD (2021) including fruit size and shape aspects. Regarding to fruit size, it shall be larger than 60mm of diameter (or 90g) while smaller sizes can be only accepted if they have a high ºBRIX level. Uniformity in size is also required, with lower range of variation allowed for higher quality classes. Defects in fruit shape are mainly considered as a cause of insufficient development and are only allowed for lower class qualities. In addition, since each variety has a characteristic of typical fruit shape, deviations from it will reduce their classification into quality classes.

Although fruit shape is among the breeding criteria in commercial apple breeding programs, breeders don't tend to breed for a particular apple shape, therefore any shape from flat oblate to tall conic are acceptable (Brown, 1960). Once a new variety is developed and ready for registration, fruit shape must be disclosed following standard descriptors (established by the Union for the Protection of New Varieties, UPOV), for it is a varietal trait used for distinctiveness, uniformity, and stability (DUS) assessments.

For scoring of key shape related characteristics, experts follow guidelines in the form of either written instructions or visual reference sketches. While the UPOV descriptors score for six differentiated shape classes and are usually used to register new varieties, other descriptors as the ones recommended by the European Cooperative Programme for Plant Genetic Resources (ECPGR) identify 13 visual classes (Szalatnay, 2006) and are usually applied for the characterization of germplasm of repositories or bank collections. The ECPGR include also metric descriptors based on Dapena et al. (2009), which defines the shape categories as the combination of two measures: the ratio between width and heigth (known as fruit shape index, FSI ratio) and the conical aspect indicated by the

ratio between the fruit width at the eye basin and stalk cavity. These authors set the numerical boundaries for FSI-derived classes at 0.75, 0.85, 0.95,1.05 and 1.15 for the flat to very long shapes, and the boundaries for the conical classes at 0.715 and 0.815 for conical to cylindrical shapes. Other works reduce the number of FSI-derived clases to three, setting the boundaries at 0.95, 1.05 and above 1.05 for the oblate spheroid, spheroid and oblong spheroid shapes, respectively (Keshavarzpour and Rashidi, 2010). Ribs and the eye bassin depth have also been included in some descriptors (Dapena et al., 2009).

So far, those parameters are measured manually and/or by comparison with the sketches provided in the guides, for ultimately sort the apples into defined classes. While these classifications are useful to describe cultivars, and despite some are based on objective measures, they do not provide quantitative or objective phenotypic evaluations of the whole fruit aspect in a way that could be used for genetic analysis purposes. A first step towards an exhaustive study of fruit shape variation requires objective mesurements of all possible fruit aspects that could best describe shape variation in diverse germplasm collections.

There are software applications that can analyze fruit images, such as Tomato Analyzer, which automatically analyzes 2D scanned images to obtain up to 37 morphological and morphometric measurements. This software was developed to analyze tomato images (Gonzalo et al., 2009) although has been successfully applied to other crops such as melon (Pereira et al., 2018), eggplant (Hurtado et al., 2013) or bell pepper (Nankar et al., 2020). However, its use and assessment efficiency in fruits like apple, with internal and external areas with shades in the calix and stalk regions, to our knowledge, has not been reported yet.

In this work we study the efficiency of Tomato Analyzer in the anlysis of apple images and provide a description of apple fruit size and shape attributes through quatitative values. For this, we used the Tomato Analyzer v3 to measure 12,920 bidimensional images of apple sections from 364 genotypes collected in three harvest seasons. We analysed the distribution of these parameters in the sample with the objective of showing their variability between accessions and years, as well as their heritability. In addition, we determined which parameters have more weight in the assignmet of the apples into visual classes, as those regularly used by breeders, evaluation officers or germplasm curators. For this, we used a Random Forest algorithm, which is a machine learning classifier that reproduces hundreds of decision trees into a predictive model representing the relationship between multiple independent variables and a dependent variable and combines them to increase the accuracy of the predictive model (Breiman, 2001). The results found with this classifier supports the use of the Dapena et al. (2009) method and provide new relevant measures that could help to refine fruit characterization.

MATERIALS AND METHODS

Fruit sampling and processing

Apples from 364 genotypes of the apple REFPOP collection were collected in three years, 2018 (143 genotypes), 2019 (276 genotypes) and 2020 (346 genotypes) being 94 genotypes common between years (**Supplementary Data 1.1**). Each genotype of the REFPOP was duplicated in a completely random block system in the IRTA fields in Gimenells (Catalunya, Spain). The REFPOP collection was managed as indicated in Jung et al., (2022).

At least three fruits representative of the fruits at the whole tree were collected at harvest maturity (assessed by iodine solution test) and stored at 4ºC until processing. For processing, clean apples were fixed with a bench vise and cut in two sections along their longitudinal axis (from stem to calyx) with a double handle knife. Per each genotype, sections of three or five fruits from the same tree were scanned into a single image with a Mustek A3 S-Series image scanner (Mustek Systems Inc., Taiwan) at 300 dpi of resolution.

Measurements obtained with Tomato Analyzer and Fruit Shape Visual Categorization

A total number of 12,920 images of fruit sections were analyzed using Tomato Analyzer (TA) version 3 software (Gonzalo et al., 2009; Rodríguez et al., 2010). After visual inspection and, when needed, manual correction of the contours determined by the software, fifteen measures providing information of fruit size and shape attributes (**Figure 1.1 and Supplementary Table 1.1**) were obtained per each fruit section. Images were, in addition, visually classified in three major classes (spheroid oblate or flat, spheroid or round, and spheroid oblong) as well as in the thirteen classes described in (Szalatnay, 2006) (**Supplementary Materials & Methods**).

Numerical descriptors and statistical analysis

Each image processed with the Tomato Analyzer software contained from six to ten fruit sections from each tree. The values used for each genotype were the mean of the values of all the apple sections of that genotype. Data were statistically analyzed using R packages. The coefficient of variation (CV) was calculated to know the variation across years, considering the total number of genotypes per year. The environmental component of the variables was analyzed in the data obtained in the collection of 94 genotypes common in the three years, which was further evaluated for normality and

homocedasticity with Shapiro Wilk's and Bartlett test, respectively. Data producing heterocedasticity were removed to perform ANOVA or Kruskal-Wallis tests (for data normally or non-normally distributed, respectively). When the null hypothesis (H_0: $\mu_{2018} = \mu_{2019} = \mu_{2020}$) was rejected, a multiple comparison test was performed to clarify differences between pairs of years using either Tukey-HDS (for the normally distributed measurements) or Dunn tests. These statistical analyses were done using *PMCMRplus* R package (Pohlert, 2014) and visualized through a *ggplot2* (Wickham, 2016). Broad sense heritability (H^2) was evaluated in the collection of 94 genotypes with the R package inti (Lozano-Isla, 2021) as the quotient between the genotypic variance (σ_p) and the phenotypic variance (σ_P) which includes the interaction between genotype x year.

As well, was analyzed the Spearman correlation (ρ) and violin plot between the measurements obtained in the whole collection of 364 genotypes were calculated and visualized through a heatmap, and the traits were clustered using *complexheatmap* R package (Gu et al., 2016). Data dispersion was measured for the measures FSII, FST PAMa, DAMa, and ECC using *seaborn* (Waskom, 2021) library in Python.

Machine Learning classifier

Random Forest

A random forest model was trained using scikit-learn library. The dataset was split in 70% for training and 30% for validation (random state set up to 80). The parameters for the random forest were 500 estimators with a max depth of 5 and 10, using our data classified to CAT-own and ECPGR respectively (Pedregosa et al., 2011). The shape variables FSII, FST, PAMa, DAMa, and ECC were selected as independent variables and the CAT-own and ECPGR categories were used as dependent variables of the genotypes.

The confusion matrix and the importance of the variables for each model were visualized by Matplotlib (Hunter, 2007) and Seaborn (Waskom, 2021) in Python.

RESULTS

Use of Tomato Analyzer software for apple shape detection

Apple images were processed with the Tomato Analyzer software. In general, the images required manual edition. The failure in the detection of the calix and stalk cavities was generalized (**Supplementary Figure 1.1**).

Size and shape attributes

The measures obtained from the Tomato Analyzer software referred to both, fruit size and shape attributes. For all traits, data were quantitative continuous; size measures were independent variables, while some of the shape descriptors were dependent as derived from the ratio between fruit size attributes such as height and width.This was the case of the fruit shape index (FSII), obtained from the ratio between fruit height and width, and of the fruit blockiness and triangle shape (PFB, DFB and FST), which were the ratios between widths at different fruit positions, and reflected conical proportions. Other shape parameters designated how well the fruit section described an ellipsoid, a circular or a rectangular (E, C, R) shape and its deviation from a circle form, term referred as eccentricity (ECC, FSIINT). Measurements in some apple areas like angles at the stalk and calyx regions with PAMa and DAMa were also obtained (**Figure 1.1**).

The number of scanned genotypes differed for the three years of evaluations, while 94 coincided in the three sample sets to allow for the estimation of environmental effects on the traits. The mean values and the coefficients of variation (CV) for the traits and years are shown in **Figure 1.2** and in **Supplementary Table 1.2**. The coefficients of variation within years ranged from 2.3% for the rectangular the area (A) values in 2018.

In general, size attributes had higher CV values than shape attributes (14.18% vs 10.87% in average, respectively).

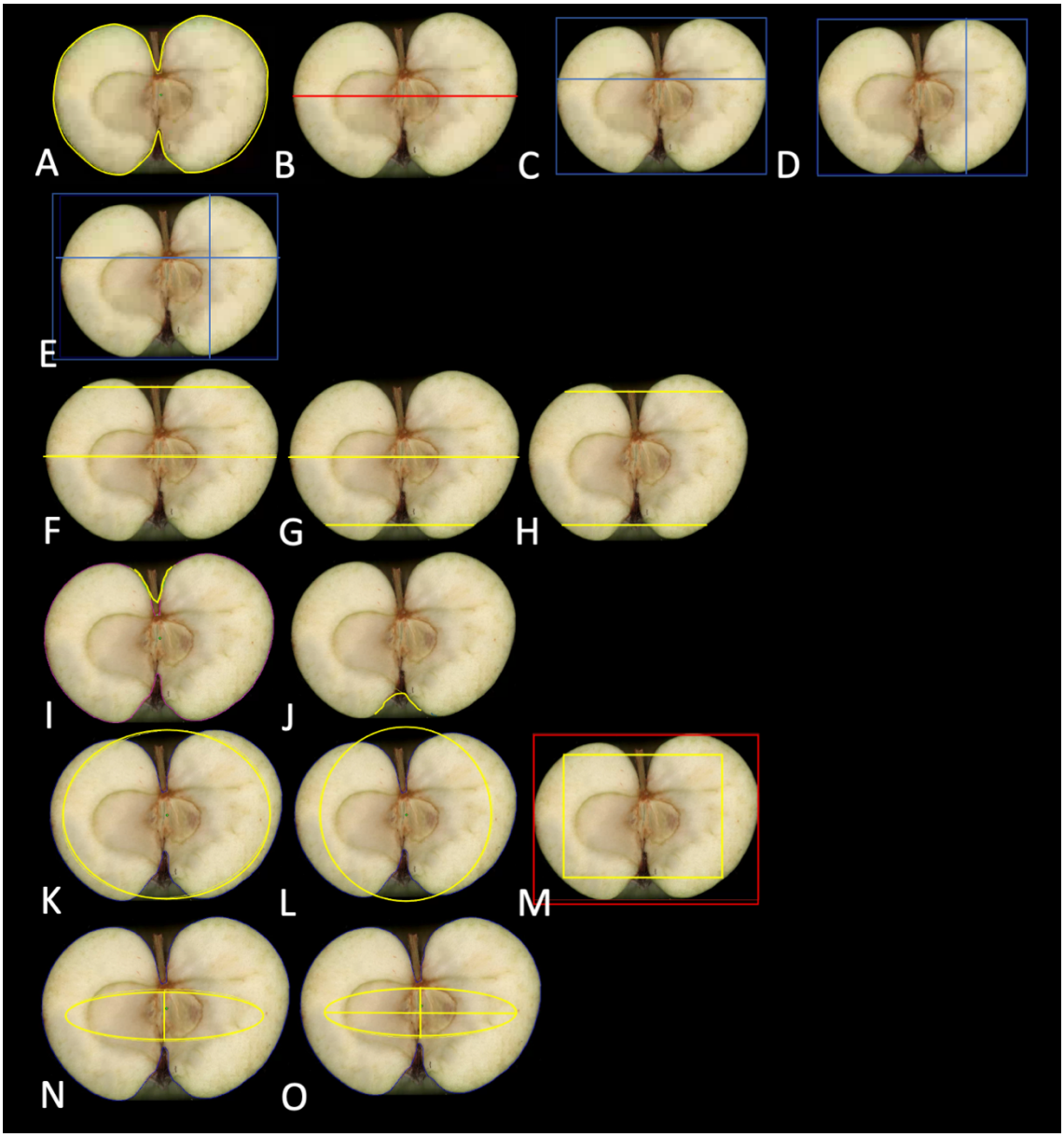


Figure 1.1. Measurements selected from Tomato Analyzer (TA) software for this study. Fifteen measures describe the apple morphological analysis in 2D of the longitudinal section fruit. Basic measurements: A. Area (A), B. Width Mid-height (WMH), C. Maximum Width (MW), D. Maximum Height (MH). Fruit Shape Index: E. Fruit shape index external I (FSII). Blockiness: F. Proximal fruit blockiness (PFB), G. Distal fruit blockiness (DFB), H. Fruit shape triangle (FST). Proximal Fruit End Shape: I. Proximal angle macro (PAMa). Distal Fruit End Shape: J. Distal angle macro (DAMa). Homogeneity: K. Ellipsoid (E), L. Circular (C), M. Rectangular (R). Internal Eccentricity: N. Eccentricity (ECC), O. Fruit shape internal (FSIINT).

A correlation analysis (Figure 1.3) clustered the traits in four main clades. Clade 1 included size measures, all with very strong correlation (ρ>0.8). Clade 2 included two of the blockiness measures PFB and FST in strong correlation (ρ =0.6); both ratios use the

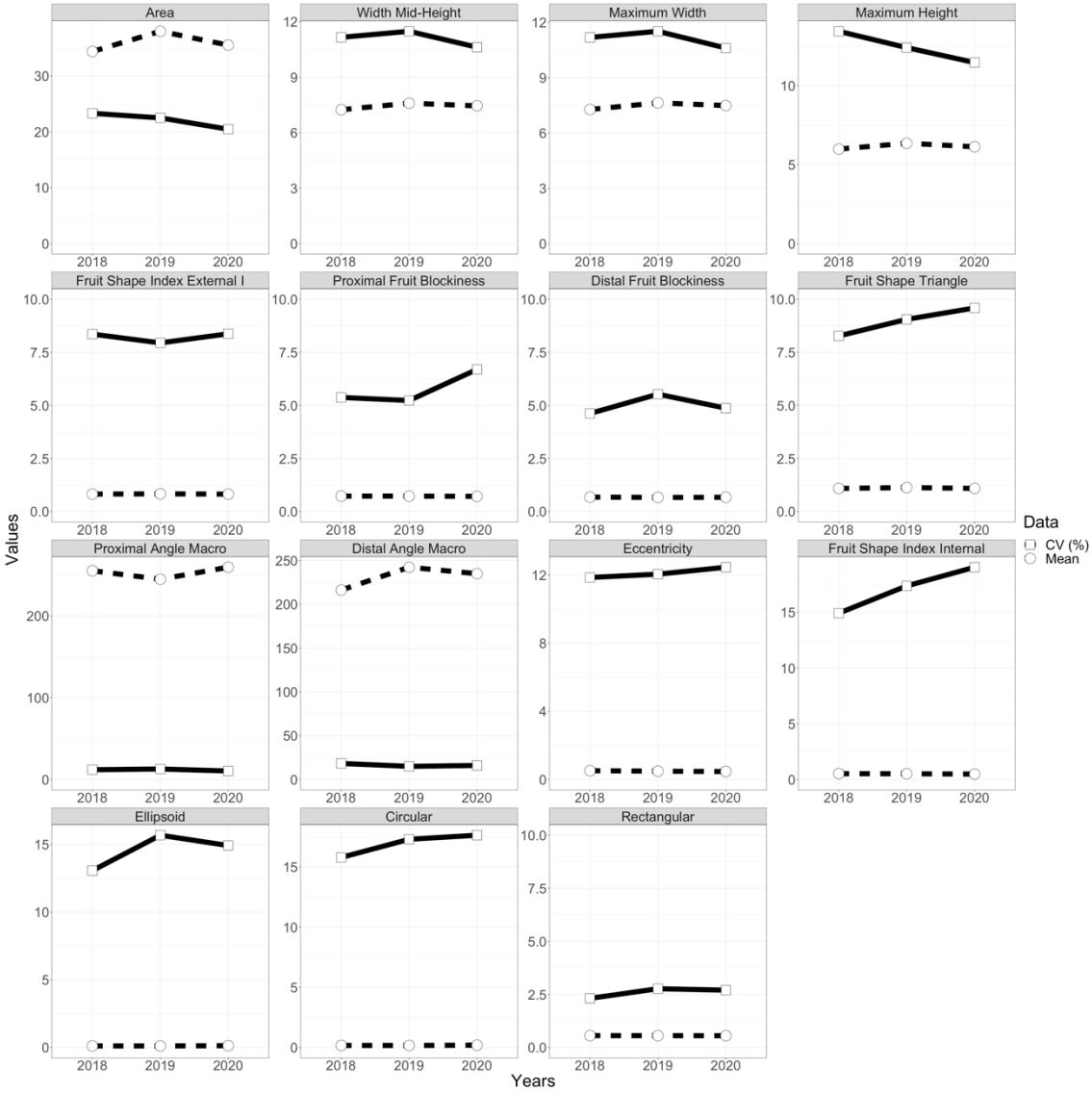


Figure 1.2. Mean and CV (coefficient of variation in percentage) values for the traits in the apple images evaluated in 2018 (143 genotypes), 2019 (276 genotypes) and 2020 (346 genotypes).

distal fruit width in their formula. Clade 3 was made of three subclades, one with the strongly correlated E and C fruit homogeneity indicators (ρ =0.8), one with the angles at the stalk and eye basin (PAMa, DAMa) which were very weakly correlated (ρ =0.1) and

the third one with rectangular shape (R) and distal fruit blockiness (DFB) in strong correlation (ρ =0.6). The Clade 4 contained the internal and external fruit shape indexes, FSII and FSIINT, strongly correlated (ρ =0.7) and the ECC, which was in moderate correlation with FSI-I (ρ=0.4) and in very stron correlatio with FSIINT (ρ=0.9). Very strong negative correlation was observed between C and ECC and FSIINT (ρ>0.8).

The violin plots show the distribution of the trait values, revealing outliers in PFB, FST, PAMa, R, DFB, and FSIINT (**Figure 1.3**).

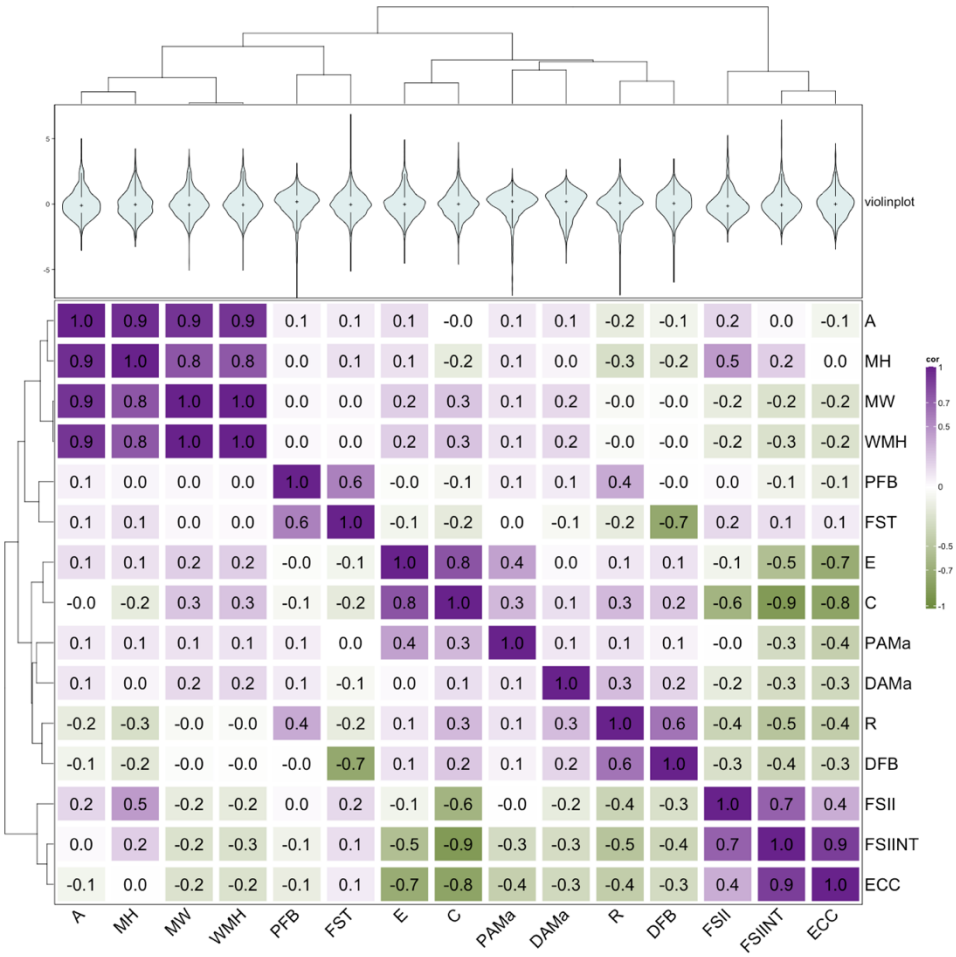


Figure 1.3. Complexheatmap of fruit measurements. Clustering of the fifteen measurements with its histogram and density violin plot, and below the correlation matrix is colored according to the Spearman's correlation coefficient using a complete dataset.

Variance between years

The 94 genotypes phenotyped in the three seasons were used to measure the environmental effect in the traits. Only the data of five traits (FST, ECC, FSIINT, E and C) presented normal distribution ($p>0.01$) after logarithmic transformation (**Supplementary Table 1.3**). Homoscedasticity failed for PFB, PAMa, DAMa and E data. For a proper analysis of their data along years, we removed the values producing the heteroscedasticity, which in most cases were produce by errors in the TA program. This meant a slight reduction of the sample size from 94 to 90 for PFB, to 91 for PAMA, to 92 for E. For the case of the DAMa, all 2020 data was discarded.

An ANOVA or a Kruskal Wallis test were conducted for normal and non-normal distributed traits, respectively. Differences between means ($p<0.001$) were observed in five out of the 15 measurements, all of them were shape attributes (FST, ECC, FSIINT, E and C). The multiple comparison test showed that in three of them, the differences occurred with the 2020 data (**Supplementary Table 1.3**). We did not observe significant variations between years in the size attributes (area, width and height).

Trait heritability ranges from 0.15 (DAMa) to 0.82 (FSII). In general, size traits had higher heritability than shape traits (0.72 *vs* 0.45 in average, respectively) (**Supplementary Table 1.4**). The shape related traits showing higher heritability after FSII were C, FSIINT and PAMa (with H^2 of 0.62, 0.57 and 0.52, respectively).

Main fruit shape descriptors

The ratio between fruit height and width (known as Fruit Shape Index, FSI) and the conical aspect of the apple are the most considered traits for the assignation of fruits into classes (Dapena et al., 2009). In our dataset, the FSII was corresponded with the Tomato Analyzer Fruit Shape External Index (FSII) and the conical aspect resembles the

blockiness measures (FST, PFB and DFB). These traits showed low CV values (from 7.90% to 9.59%), indicating moderate variation in the samples evaluated.

The FSII reached values from 0.67 to 1.14 (**Figure 1.4A**, **Supplementary Table 2 and 3**), with a mean of 0.825 and an average CV of 8.22%. 'Gros Api' and 'Belle Flavoise', described as flat varieties in cultivar databases, were among the ones with lower FSII together with others like 'Grenadier' and 'Szaszpap Alma' described as broad globose conical (http://www.nationalfruitcollection.org.uk/index.php). In contrast, varieties such as 'Skovfoged', 'Giambun', 'Rosmarina blanca', 'Boordin Negal' and 'Maglemer' had FSII ratios higher than 1.1; these varieties were described as ellipsoid conical, narrow conical, truncate conical and conical (http://www.nationalfruitcollection.org.uk/index.php).

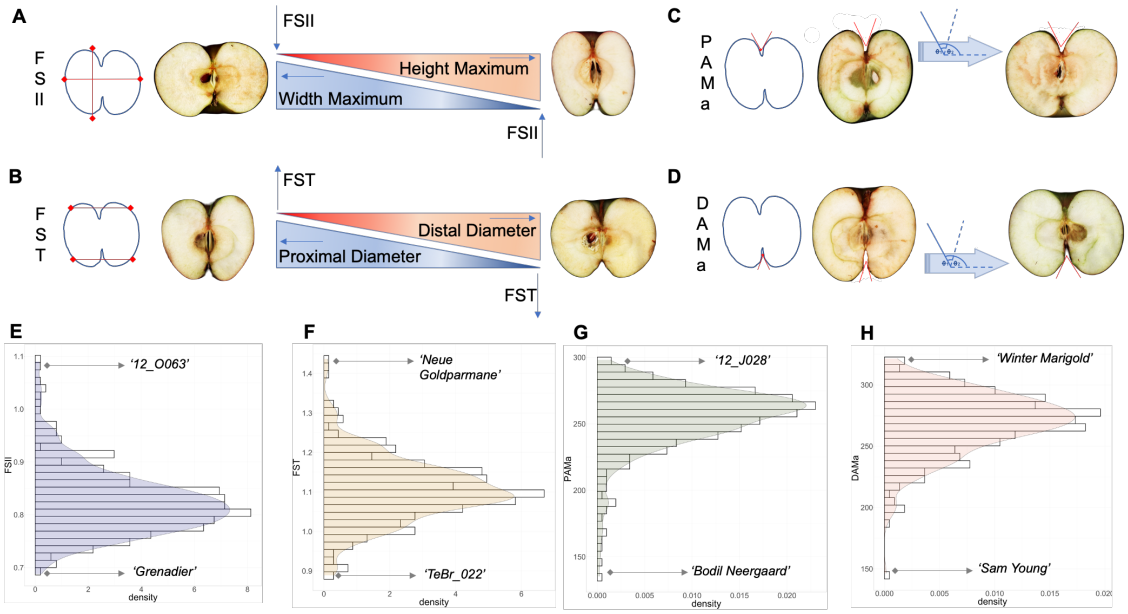


Figure 1.4. Depiction of apple fruit shape, showing fruit parts from low-scoring and high-scoring genotypes and their density for each measurement. Proximal fruit (stem), distal fruit (calyx) **A** and **E**. FSII, **B** and **F**. FST, **C** and **G**. PAMa, **D** and **H**. DAMa.

The FST was the ratio between the width at the stalk cavity (proximal) and eye basin (distal) shoulders (**Figure 1.1 and Figure 1.4B**). The width at the proximal or at the distal

shoulders were also used in the PFB and in the DFB, respectively. In consequence FST had strong correlation with PFB (ρ=0.6) and strong negative correlation with DFB (ρ=-0.7). The average of the FST value in the dataset was 1.1, with an average CV of 9.3% when considering the three-year data. Some varieties showed contrasting values between years (**Supplementary Table 1.2**). This was the case of 'Reinette d'Anthezieux' (FST_{2019}=1.743; FST_{2020}=1.052) and 'Reinete Sanguine du Rhin' (FST_{2019}=1.155; FST_{2020}=0.643), for example. A revision of the images revealed a detection error of the TA program, which does not recognize well the shoulder limits in asymmetric apple sections (**Supplementary Figure 1.1**). The varieties 'Neue Goldparmane' and 'Priscilla' showed high FST values in two years (FST_{2019}=1.455, FST_{2020}=1.419 and FST_{2019}=1.271 and FST_{2020}=1.358, respectively), in contrast to 'Winesap' that had low FST in both 2019 and 2020 evaluations (0.922 and 0.909, respectively). Despite having contrasting FST values, all three varieties are described as conical in apple collection databases.

Supervised machine learning

The apples in 765 of the scanner images (each containing between 12 to 20 apple sections) were annotated following two classifications: 1) a simple one considering only three classes (spheroid oblate, spheroid and spheroid oblong) that we called CAT-own, and 2) the ECPGR catalog (**Supplementary Data 1.2**).

Figure 1.5 shows the data dispersion using the whole scanned images dataset and five shape parameters (FSII, FST, PAMa, DAMa, ECC) selected as important variables that could describe the shape. Between theses shape parameters, the FSII measure separates almost all data according to their classes by CAT-own, while some of the ECPGR classes overlapped; this is the case of the broad-globose-conical and flat globose classes, globose conical and globose classes.

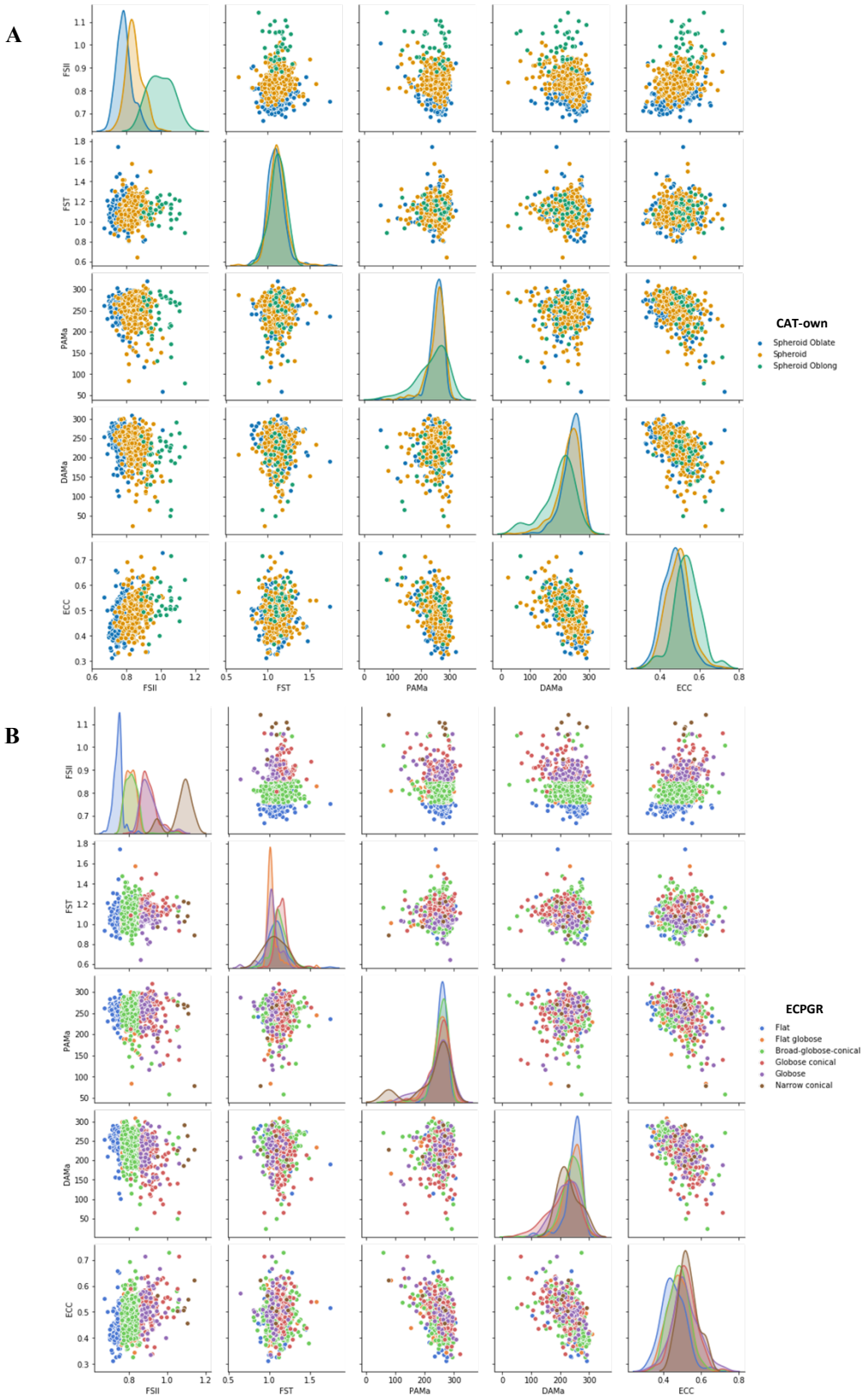


Figure 1.5. Scatterplot matrix of shape data in the fruit shape categories. **A**, CAT-visual category, being the class as blue is spheroid oblate, orange is spheroid and green is spheroid oblong. **B**, ECPGR criterion was categorized in six shapes, blue is flat, orange is flat globose, green is Broad-globose-conical, red is globose conical, purple is globose and brown is narrow conical.

We used a supervised machine learning classifier to identify the traits relevants for fruit classification as well as their weight into the classification. The random forest algorithm was performed using data from the three years. For the data classified by ECPGR the narrow conical class was excluded due a low number of data in this class. Therefore, we ran one classifier per dataset (CAT-own and ECPGR). In the CAT-own, the model obtained a high accuracy of 0.90 and a f1-score between 0.82 and 0.92 in the classes. In terms of the number of data assigned correctly per class, the spheroid oblong performed worst. In the ECPGR classes, the accuracy is 0.9, and an f1-score of 0.71 to 0.98 in the classes (**Table 1.1**).

According to random forest feature importance, the variable with more relevance for each model was the FSII measure (**Figure 1.6A**). In addition, in the ECPGR classification the FST (a descriptor of the fruit conicity) and the DAMa (a descriptor of the eye basin) follow FSI in relevance (**Figure 1.6B**). The confusion matrix allowed the visualization of the performance of each model showed that in the CAT-own model 10 of the 102 spheroid fruits were predicted as spheroid oblate, and that nine of the 118 spheroid oblate were predicted as spheroid with the parameters selected as relevan. The number of samples with spheroid olong shape was reduced; three out of 10 were wrongly predicted as spheroids (**Figure 1.6C**).

In the ECPGR model, with five categories in our analysis, only seven of the broad-globose-conical class were wrongly predicted (six as globose and one as globose conical).

Table 1.1. Classification report of random forest. Two models predictive, CAT-own and ECPGR.

Model	Categories	Precision	Recall	F1-score	Support
CAT-own	Spheroid	0,88	0,9	0,89	102
	Spheroid Oblate	0,92	0,92	0,92	118
	Spheroid Oblong	1	0,7	0,82	10
	Accuracy			0,9	230
	Macro avg	0,93	0,84	0,88	230
	Weighted avg	0,91	0,9	0,9	230
ECPGR	Broad-globose-conical	0,9	0,94	0,92	110
	Flat	1	0,97	0,98	32
	Flat globose	0,75	0,67	0,71	27
	Globose	0,85	0,85	0,85	13
	Globose conical	0,93	0,93	0,93	46
	Accuracy			0,9	228
	Macro avg	0,89	0,87	0,88	228
	Weighted avg	0,9	0,9	0,9	228

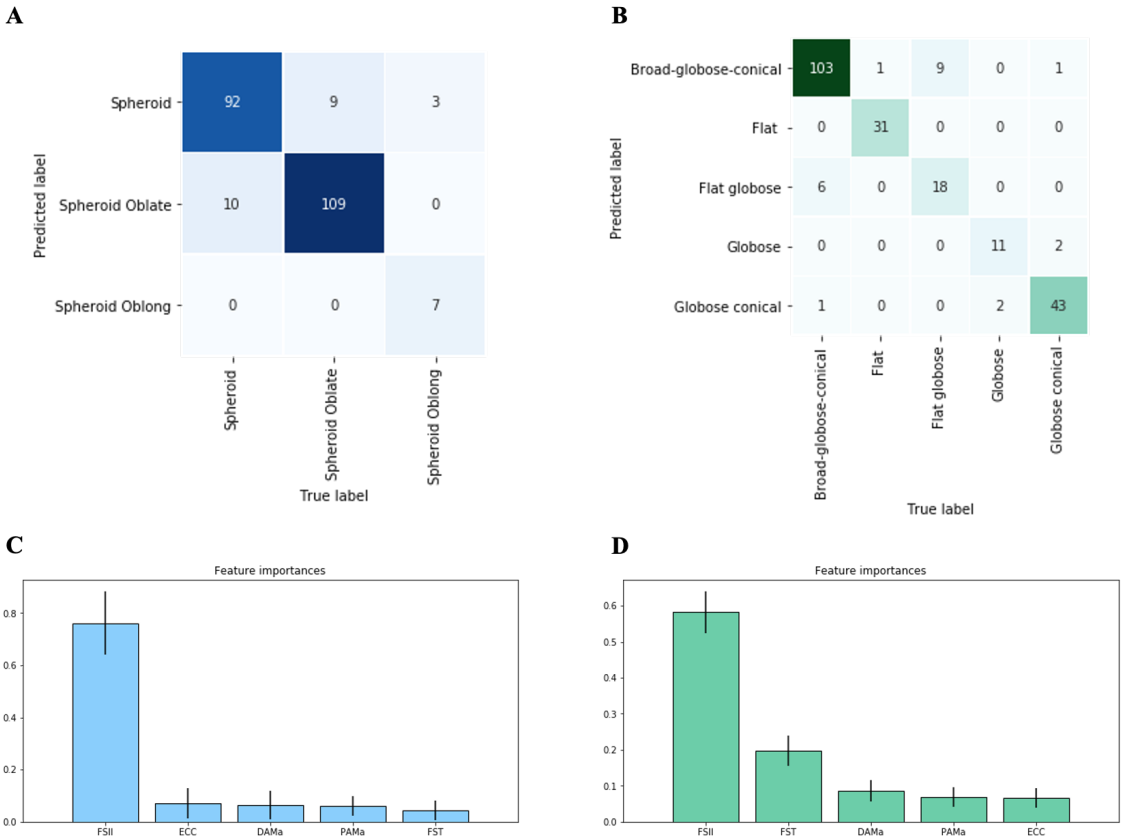


Figure 1.6. Random forest results. Variables importances for predictive model **A**, CAT-own. **B**, for the ECPGR. Confusion matrix, x axis as true label and y axis as predicted label. **C**, CAT-own and **D**, ECPGR.

The prediction efficiency for the flat apples was also very high, with only one missmatch out of 32. Flat globose shape was the worst predicted, one third of the times (nine out of 27) were wrongly asigned as broad globlose conical. (**Figure 1.6D**).

DISCUSSION

While the guidelines dictated to classify apples according to their shape are mainly based on the ratio between height and diameter and the ratio between the width of the shoulders at both ends of the fruit (UPOV, ECPGR), there are other attributes that, although less relevant in the perception of shape, may also contribute to the overall perception of the shape of the fruit. For this reason, we conducted an high-throughput phenotyping assay that provided data for 15 fruit size and shape attributes.

To easily acquire the phenotypic data, we cut apples from genotypes included in a broad germplasm collection representative of the genetic variability in European germplasm (the apple REFPOP), scanned the sections and processed the 2D images with a software designed to evaluate shape in tomato, the Tomato Analyzer software. Despite its applicability for the analysis of fruits of different species and even leaves as described in multiple works (Gonzalo et al., 2009; Nankar et al., 2020; Pereira et al., 2021; Sierra-Orozco et al., 2021), its use to evaluate apple fruit sections has been a challenge as each fruit section had to be adjusted manually to correct identification. However, as advantage, it provides numerous size and shape related measurements.

In apple, several works have used image analysis tools to classify cultivars with digital images of seeds (Sau et al., 2019), flower buds using deep learning in image recognition tasks such as a data enrichment network (Xia et al., 2021) or to characterize the internal browning in apples using X-ray computed tomography analysis of 3D images (Chigwaya et al., 2021).

In the UPOV varietal guide for cultivar characterization, there are different descriptors of apple tree morphology from flowering, production cycle and especially the fruit quality. Shape descriptors are based on FSI rations as well as similitude with sketches and reference cultivars. In other guides apple fruit shape attributes are characterized by measuring ratios in addition to the FSI as that from the stalk cavity and the eye basin (Szalatnay, 2006; Dapena et al., 2009). Here we obtained similar measurements such as the FSII (height/diameter) and the FST (average stalk cavity/eye basin diameter) as well as other more difficult to evaluate in whole fruits or with manual tools as a caliper. Although asymmetry was observed, it could not be measured with Tomato Analyzer. Fruit symmetry is an important quality parameter, usually associated to development aspects although with a varietal (genetic) component. Some varieties are characterized by showing asymmetrical shape, like 'Brabant Bellefleur' see description in (https://www.fruitid.com/#view/487). Lopsidedness is associated with irregular seed weight, and, in a minor way, to the number of seeds in the five carpel derived sectors (Brault and de Oliveira, 1995; Drazeta et al., 1999). A reason for underdeveloped seeds could be found in an inadequate partial pollination (Matsumoto et al., 2012). The genetic diversity and the field design of the REFPOP discard the pollination effect in the apples analyzed. All trees were in the same orchard and cultivated under the same management; in addition, the apples were selected as representative of the fruits in the tree and the data used for the analyses was the mean of 12 to 20 fruit sections per genotype. These should have captured the fruit traits characteristic of the variety while reducing the environmental effect of fruit asymmetry in the data.

For some attributes, the variance between years showed considerable fluctuations, probably because of environmental reasons. Monthly average temperatures during the

fruiting period (from April 1st to October 30th) were very similar in the three years (**Supplementary Table 1.4**), contrasting with the average maximum and minimum daily temperatures and with the daily accumulated precipitation, which showed some important deviations. Suggesting a possible water or heat stress that probably was mitigated by the field irrigation system.

Apple size and shape are inherited independently. Variability and heritability of apple shape has been reported in few breeding studies using a reduced number of varieties. Our results are in agreement with Crane and Lawrence (1933), who found that variation of FSII between years was not wide. Interestingly Brown (1960) reported that mean FSII of a progeny is approximately midway between the two parents. Both results indicate a high genetic component of FSII measure and, therefore, the usefulness of this parameter in breeding.

In addition to the development of algorithm-based programs, artificial intelligence methods are important tools that allow the identification and prediction of large scale interpretations and were implemented in various fields of genetics and genomics (Libbrecht and Noble, 2015). For example, the use of machine learning was implemented from pre-harvest to post-harvest in agriculture (Meshram et al., 2021). In addition, it is used to measure by advanced electronic devices (Biffi et al., 2021; Gongal et al., 2018; Häni et al., 2020; Tsoulias et al., 2020).

As Random Forest relays in the construction of multiple decision trees, it retains their advantages while using grouped samples, random variable subsets to achieve better results, and handles missing values. As well as allow to use of several types of variables (continuous, binary, and categorical) and it is suitable for modeling multidimensional data (Qi, 2012).

Currently, machine learning and deep learning were implemented in agriculture, giving several tools for characterizing and selecting interesting genetic material in a breeding program (Danckaers et al., 2017; Meshram et al., 2021; Zhang et al., 2021). Despite of the interest in the automatization of the phenotyping measurement, for apple there is not a specific free software for the correct recognition of the fruit morphology; nonetheless, the TA software brough interesting measures. Recently phenomics analysis including pipelines for 2D image segmentations have been published for strawberry and other fruits (Zingaretti et al., 2021). However, this pipeline needs code tuning for its application in apple. Also, 3D reconstructions have been applied in other fruit shape studies (Wang and Chen, 2020; He et al., 2017). The Zingaretti et al. (2021) needs code tuning for its application in apple. Fruit shape has been studied in eggplant (Mangino et al., 2021), tomato (Gonzalo et al., 2009).

Random Forest has been used to classify crop related traits. For example in the rice the surpervised classification was used for the analysis of local Panamanian rice crpos using plant phenology and Near-Infrared (NIR) traits (Sánchez-Galán et al., 2021). Moradi et a. (2021) used the approache to predict growth regions for *Moringa peregrina* (an tree species that contributes to the restoration of fragile ecosystems. In wheat, Zhang et al. (2021) used it to predict winter wheat leaf water content. In tomato it was used to find important variables to predict tomato yields by aerial vehicle imagery (Tatsumi et al., 2021). In our study, the Random Forest algorithm was able to predict most of the shapes of the two classifications with high precision, and extreme classes were not confounded. The Cat-own classification method assigned the apples into three classes as they were perceived by the human criteria. In the case of the ECPGR classification, the assignation was according to the comparisons with sketches that also contemplate conicity. Our

results suggest that the FSII and FST values obtained with the Tomato Analyzer can be well used for the automatic classification of samples into classes.

CONCLUSION

In this work, we did a high throughput phenotyping assay of apple fruit morphology using images in 2D, analyzing the variability from germplasm in years and measurements. The software used (Tomato Analyzer) produced accurate measurements, although previous manual modification was needed for most of the images, limiting its use in the high-throughput phenotyping of apple samples. Some of the traits evaluated had high heritability, indicating an important genetic component in their determination. Among the most informative traits determining fruit shape, we found FSII and FST, followed by DAMa, ECC and PAMa. The heritability of these traits encourages their use in genomic studies.

REFERENCES

1. Biffi, L. J., Mitishita, E. A., Liesenberg, V., Centeno, J. A. S., Schimalski, M. B., & Rufato, L. (2021). Evaluating the performance of a semi-automatic apple fruit detection in a high-density orchard system using low-cost digital RGB imaging sensor. *Boletim de Ciências Geodésicas*, *27*.
2. Brault, A. M., & de Oliveira, D. (1995). Seed Number and an Asymmetry Index of McIntosh' Apples. *HortScience*, *30*(1), 44-46.
3. Breiman, L. (2001). Random forests. *Machine learning*, *45*(1), 5-32.
4. Brown, A. G. (1960). The inheritance of shape, size and season of ripening in progenies of the cultivated apple. *Euphytica*, *9*, 327-337.
5. Chigwaya, K., du Plessis, A., Viljoen, D. W., Crouch, I. J., & Crouch, E. M. (2021). Use of X-ray computed tomography and 3D image analysis to characterize internal browning in 'Fuji'apples after exposure to CO2 stress. *Scientia Horticulturae*, *277*, 109840.
6. Crane, M. B., & Lawrence, W. J. C. (1933). Genetical studies in cultivated apples. *Journal of Genetics*, 28(2), 265-296.

7. Danckaers, F., Huysmans, T., Dael, M. V., Verboven, P., Nicolaï, B., & Sijbers, J. (2017). Building 3D statistical shape models of horticultural products. *Food and bioprocess technology*, *10*(11), 2100-2112.

8. Dapena, E., & Blázquez, M. (2009). Descripción de las variedades de manzana de la DOP Sidra de Asturias. *SERIDA, Asturias*.

9. Drazeta, L., Lang, S., Hall, A., Volz, R., & Jameson, P. E. (1999, February). Seed set and the development of fruit shape in apple. In *Proceedings of a Seed Symposium. Massey University, Palmerston North, New Zealand* (Vol. 12, pp. 99-101).

10. Gongal, A., Karkee, M., & Amatya, S. (2018). Apple fruit size estimation using a 3D machine vision system. *Information Processing in Agriculture*, *5*(4), 498-503.

11. Gonzalo, M. J., Brewer, M. T., Anderson, C., Sullivan, D., Gray, S., & van der Knaap, E. (2009). Tomato fruit shape analysis using morphometric and morphology attributes implemented in Tomato Analyzer software program. *Journal of the American Society for Horticultural Science*, *134*(1), 77-87.

12. Gu, Z., Eils, R., & Schlesner, M. (2016). Complex heatmaps reveal patterns and correlations in multidimensional genomic data. *Bioinformatics*, *32*(18), 2847-2849.

13. Häni, N., Roy, P., & Isler, V. (2020). A comparative study of fruit detection and counting methods for yield mapping in apple orchards. *Journal of Field Robotics*, *37*(2), 263-282.

14. He, J. Q., Harrison, R. J., & Li, B. (2017). A novel 3D imaging system for strawberry phenotyping. *Plant Methods*, *13*(1), 1-8.

15. Hunter, J. D. (2007). Matplotlib: A 2D graphics environment. *Computing in science & engineering*, *9*(03), 90-95.

16. Hurtado, M., Vilanova, S., Plazas, M., Gramazio, P., Herraiz, F. J., Andújar, I., & Prohens, J. (2013). Phenomics of fruit shape in eggplant (Solanum melongena L.) using Tomato Analyzer software. *Scientia Horticulturae*, *164*, 625-632.

17. Jung, M., Keller, B., Roth, M., Aranzana, M. J., Auwerkerken, A., Guerra, W., ... & Patocchi, A. (2022). Genetic architecture and genomic predictive ability of apple quantitative traits across environments. *Horticulture research*, *9*.

18. Keshavarzpour, F., & Rashidi, M. (2010). Classification of apple size and shape based on mass and outer dimensions. *Am.-Eurasian J. Agric. Environ. Sci*, *9*(6), 618-621.

19. Libbrecht, M. W., & Noble, W. S. (2015). Machine learning applications in genetics and genomics. *Nature Reviews Genetics*, *16*(6), 321-332.

20. Lozano-Isla, F. (2021). Inti: Tools and Statistical Procedures in Plant Science. *R Package Version 0.1*, *3*.

21. Mangino, G., Vilanova, S., Plazas, M., Prohens, J., & Gramazio, P. (2021). Fruit shape morphometric analysis and QTL detection in a set of eggplant introgression lines. *Scientia Horticulturae*, *282*, 110006.

22. Matsumoto, S., Soejima, J., & Maejima, T. (2012). Influence of repeated pollination on seed number and fruit shape of 'Fuji'apples. *Scientia horticulturae*, *137*, 131-137.

23. Meshram, V., Patil, K., Meshram, V., Hanchate, D., & Ramkteke, S. D. (2021). Machine learning in agriculture domain: A state-of-art survey. *Artificial Intelligence in the Life Sciences*, *1*, 100010.

24. Meshram, V., Patil, K., Meshram, V., Hanchate, D., & Ramkteke, S. D. (2021). Machine learning in agriculture domain: A state-of-art survey. *Artificial Intelligence in the Life Sciences*, *1*, 100010.

25. Moradi, E., Abdolshahnejad, M., Hassangavyar, M. B., Ghoohestani, G., da Silva, A. M., Khosravi, H., & Cerdà, A. (2021). Machine learning approach to predict susceptible growth regions of Moringa peregrina (Forssk). *Ecological Informatics*, *62*, 101267.

26. Nankar, A. N., Tringovska, I., Grozeva, S., Todorova, V., & Kostova, D. (2020). Application of high-throughput phenotyping tool Tomato Analyzer to characterize Balkan Capsicum fruit diversity. *Scientia Horticulturae*, *260*, 108862.

27. OECD. (2021). Organisation for Economic Co-operation and Development. https://www.oecd.org/

28. Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., ... & Duchesnay, E. (2011). " Scikit-learn: Machine Learning in Python," Journal of Machine Learning Research, vol. 12, p.

29. Pereira, L., Ruggieri, V., Pérez, S., Alexiou, K. G., Fernández, M., Jahrmann, T., ... & Garcia-Mas, J. (2018). QTL mapping of melon fruit quality traits using a high-density GBS-based genetic map. *BMC plant biology*, *18*(1), 1-17.

30. Pereira, L., Zhang, L., Sapkota, M., Ramos, A., Razifard, H., Caicedo, A. L., & van Der Knaap, E. (2021). Unraveling the genetics of tomato fruit weight during crop domestication and diversification. *Theoretical and Applied Genetics*, *134*(10), 3363-3378.

31. Pohlert, T. (2014). The pairwise multiple comparison of mean ranks package (PMCMR). *R package*, *27*(2019), 9.

32. Qi, Y. (2012). Random forest for bioinformatics. In *Ensemble machine learning* (pp. 307-323). Springer, Boston, MA.

33. Rodríguez, G. R., Moyseenko, J. B., Robbins, M. D., Morejón, N. H., Francis, D. M., & van der Knaap, E. (2010). Tomato Analyzer: a useful software application to collect accurate and detailed morphological and colorimetric data from two-dimensional objects. *JoVE (Journal of Visualized Experiments)*, (37), e1856.

34. Sánchez-Galán, J. E., Barranco, F. R., Reyes, J. S., Quirós-McIntire, E. I., Jiménez, J. U., & Fábrega, J. R. (2019). Using Supervised Classification Methods for the Analysis of Multi-spectral Signatures of Rice Varieties in Panama.

35. Sau, S., Ucchesu, M., D'hallewin, G., & Bacchetta, G. (2019). Potential use of seed morpho-colourimetric analysis for Sardinian apple cultivar characterisation. *Computers and Electronics in Agriculture*, *162*, 373-379.

36. Sierra-Orozco, E., Shekasteband, R., Illa-Berenguer, E., Snouffer, A., van der Knaap, E., Lee, T. G., & Hutton, S. F. (2021). Identification and characterization of GLOBE, a major gene controlling fruit shape and impacting fruit size and marketability in tomato. *Horticulture research*, *8*.

37. Szalatnay, D., & Bauermeister, R. (2006). Obst-Deskriptoren NAP. *Stutz Druck AG*, *8820*.

38. Tatsumi, K., Igarashi, N., & Mengxue, X. (2021). Prediction of plant-level tomato biomass and yield using machine learning with unmanned aerial vehicle imagery. *Plant methods*, *17*(1), 1-17.

39. Tsoulias, N., Paraforos, D. S., Xanthopoulos, G., & Zude-Sasse, M. (2020). Apple shape detection based on geometric and radiometric features using a LiDAR laser scanner. *Remote Sensing*, *12*(15), 2481.

40. Wang, Y., & Chen, Y. (2020). Fruit morphological measurement based on three-dimensional reconstruction. *Agronomy*, *10*(4), 455.

41. Waskom, M. L. (2021). Seaborn: statistical data visualization. *Journal of Open Source Software*, *6*(60), 3021.

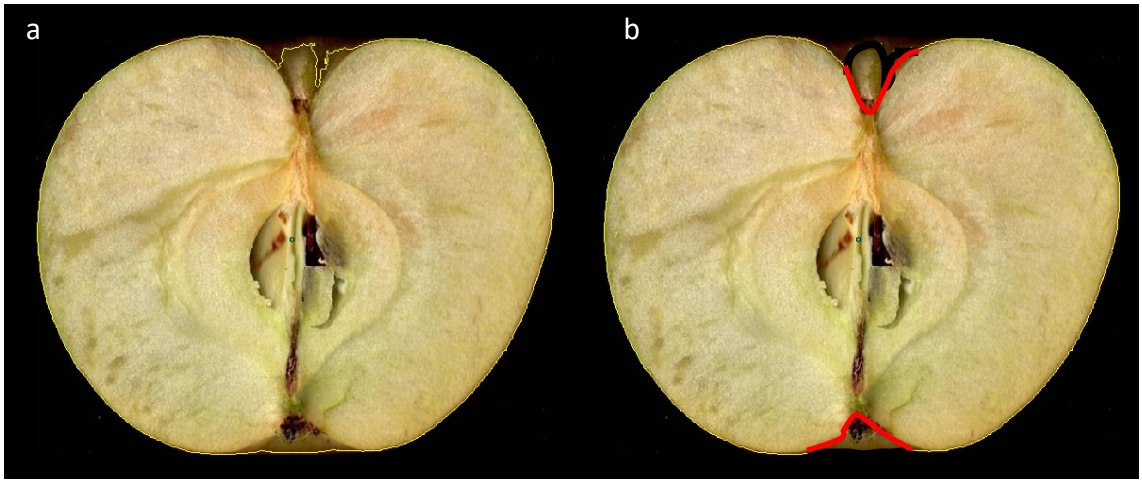
42. Wickham, H., Chang, W., & Wickham, M. H. (2016). Package 'ggplot2'. *Create elegant data visualisations using the grammar of graphics. Version*, *2*(1), 1-189.

43. Xia, X., Chai, X., Zhang, N., & Sun, T. (2021). Visual classification of apple bud-types via attention-guided data enrichment network. *Computers and Electronics in Agriculture*, *191*, 106504.

44. Zhang, J., Zhang, W., Xiong, S., Song, Z., Tian, W., Shi, L., & Ma, X. (2021). Comparison of new hyperspectral index and machine learning models for prediction of winter wheat leaf water content. *Plant Methods*, *17*(1), 1-14.

45. Zingaretti, L. M., Monfort, A., & Pérez-Enciso, M. (2021). Automatic fruit morphology phenome and genetic analysis: An application in the octoploid strawberry. *Plant Phenomics*, *2021*.

SUPPLEMENTARY MATERIAL



Supplementary Figure 1.1. Measurement errors Tomato Analyzer software version 3, using images of a longitudinal section of the apple. **a**, represents the incorrect delimitation in the proximal and distal part of the fruit. **b**, represents the manual correction marked in red. **c**, not recognizing shoulder boundaries in asymmetrical block sections.

SUPPLEMENTARY MATERIAL & METHODS

Visual parameter to classify the apple fruit shape with own category (CAT-own) and depend on the shape tendency global by tree.

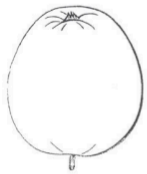
Spheroid oblate or flat

Spheroid or round

Spheroid Oblong

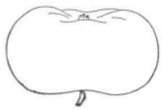
Visual parameter to classify the apple fruit shape based on Szalatnay 2006.

1= Globose

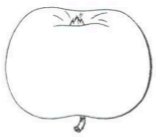


2= Globose conical

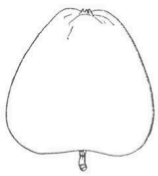
3= Broad-globose-conical



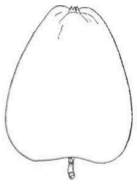
4= Flat



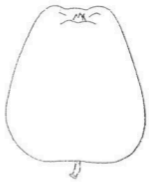
5= Flat globose



6= Conical



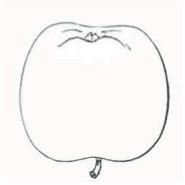
7= Narrow conical



8= Truncate conical

9= Ellipsoid

10= Ellipsoid conical



11= Oblong

12= Oblong conical

13= Waisted fruit

SUPPLEMENTARIES TABLES

Supplementary Table 1.1. List of the attributes evaluated with the Tomato Analyzer Software

Attributes		Measurements	Acronym	Measuring*
Size Attributes		Area	A	Space bounded by the perimeter (mm2)
		Width Mid-height	WMH	The width measured at ½ of the fruit's height (mm)
		Maximum Width	MW	The maximum horizontal distance of the fruit (mm)
		Maximum Height	MH	The maximum vertical distance of the fruit (mm)
Shape Attributes	**Fruit Shape Index**	Fruit Shape Index external I	FSII	The ratio of the Maximum Height to Maximum Width.
	Blockiness	Proximal Fruit Blockiness	PFB	The ratio of the width at the Upper Blockiness Position (U) to Width Mid-height
		Distal Fruit Blockiness	DFB	The ratio of the width at the Lower Blockiness Position (L) to Width Mid-height
		Fruit Shape Triangle	FST	The ratio of the width at the Upper Blockiness Position (U) to the width at the Lower Blockiness Position (L)
	Fruit End Shape	Proximal Angle Macro	PAMa	The angle between best-fit lines drawn through the fruit perimeter on either side of the proximal end point. The Macro Distance setting determines the percentage of the perimeter from the proximal end point at which to center the linear regression points used to find the best-fit line. The points comprising 5% of the perimeter on either side of that center point are used in the regression.
		Distal Angle Macro	DAMa	The angle between best-fit lines drawn through the fruit perimeter on either side of the distal end point. The Macro Distance setting determines the percentage of the perimeter from the distal end point at which to center the linear regression points used to find the best-fit line. The points comprising 5% of the perimeter on either side of that center point are used in the regression.
	Internal Eccentricity	Eccentricity	ECC	The ratio of the height of the internal ellipse to the Maximum Height.
		FSI Internal	FSIINT	The ratio of the internal ellipse's height to its width.
	Homogeneity	Ellipsoid	E	The ratio of the error resulting from a best-fit ellipse to the area of the fruit. Error is the average magnitude of residuals (Res) along the fruit's perimeter, divided by the length of the major (longer) axis of the ellipse. Smaller values indicate that the fruit is more ellipsoid.
		Circular	C	The ratio of the error resulting from a best-fit circle to the area of the fruit. Error is the average magnitude of residuals (Res) along the fruit's perimeter, divided by the radius of the circle. Smaller values indicate that the fruit is more circular.
		Rectangular	R	The ratio of the area of the rectangle bounding the fruit to the area of the rectangle bounded by the fruit.

*The descriptions of each measure can be found in detail in Rodriguez et al. (2010)

Supplementary Table 1.2. Descriptive statistics in the dataset of 2018, 2019, and 2020 evaluated with measures obtained with Tomato Analyzer software.

Measurements	Code	Number of genotypes	Years evaluated	Mean	Range		CV (%)
					Min	Max	
Basic Measurements							
Area (cm^2)	A	143	2018	34,378	21,475	63,715	23,30
		276	2019	37,958	12,994	71,087	22,50
		346	2020	35,521	18,795	61,47	20,47
Width Mid-height (cm)	WMH	143	2018	7,24	5,709	10,33	11,20
		276	2019	7,594	3,793	10,425	11,50
		346	2020	7,449	5,556	10,209	10,62
Maximum Width (cm)	MW	143	2018	7,279	5,731	10,368	11,20
		276	2019	7,637	3,815	10,474	11,50
		346	2020	7,492	5,588	10,281	10,61
Maximum Height (cm)	MH	143	2018	5,978	4,413	8,602	13,40
		276	2019	6,347	4,272	8,876	12,40
		346	2020	6,122	4,208	8,092	11,45
Fruit Shape Index							
Fruit Shape Index external I	FSII	143	2018	0,822	0,669	1,059	8,40
		276	2019	0,833	0,688	1,141	7,90
		346	2020	0,820	0,7	1,106	8,37
Blockiness							
Proximal Fruit Blockiness	PFB	143	2018	0,727	0,583	0,811	5,40
		276	2019	0,728	0,554	0,817	5,20
		346	2020	0,716	0,437	0,832	6,70
Distal Fruit Blockiness	DFB	143	2018	0,68	0,57	0,766	4,60
		276	2019	0,665	0,488	0,738	5,50
		346	2020	0,671	0,56	0,749	4,87
Fruit Shape Triangle	FST	143	2018	1,089	0,876	1,504	8,30
		276	2019	1,128	0,834	1,743	9,10
		346	2020	1,091	0,643	1,478	9,59
Proximal Fruit End Shape							
Proximal Angle Macro	PAMa	143	2018	255,254	84,515	319,335	12,00
		276	2019	244,908	58,15	296,08	12,90
		346	2020	259,542	123,447	308,561	10,42
Distal Fruit End Shape							
Distal Angle Macro	DAMa	143	2018	216,248	65,204	288,009	18,40
		276	2019	242,26	23,15	308,38	15,00
		346	2020	234,910	50,156	301,786	16,07
Internal Eccentricity							
Eccentricity	ECC	143	2018	0,517	0,338	0,713	11,90
		276	2019	0,49	0,337	0,728	12,00
		346	2020	0,468	0,314	0,715	12,45
Fruit Shape Index Internal	FSIINT	143	2018	0,526	0,346	0,731	14,90
		276	2019	0,517	0,335	0,927	17,40
		346	2020	0,489	0,281	1,034	19,04
Homogeneity							
Ellipsoid	E	143	2018	0,115	0,082	0,171	13,10
		276	2019	0,112	0,041	0,167	15,70
		346	2020	0,133	0,071	0,211	14,92
Circular	C	143	2018	0,169	0,103	0,238	15,80
		276	2019	0,166	0,047	0,233	17,30
		346	2020	0,193	0,087	0,314	17,63
Rectangular	R	143	2018	0,561	0,511	0,587	2,30
		276	2019	0,557	0,497	0,598	2,80
		346	2020	0,554	0,461	0,599	2,70

Supplementary Table 1.3. Analysis of variance between years evaluated in 94 genotypes common with measures described in Tomato Analyzer. The level of significance expressed is not significant (ns) *P < 0.05, **P < 0.01, ***P < 0.001.

		Before		After (Data logarithmized)			
Measurements	**Acronyms**	**Shapiro-Wilk's test (p-value)**	**Bartlett's test (p-value)**	**Shapiro-Wilk's test (p-value)**	**Bartlett's test (p-value)**	**ANOVA / Kruskal-Wallis (p-value)**	**Test multiple comparison (p-value)**
Basic Measurements							
Area	A	1,39E-08	0,4538	0,0008464	0,2918	0,1773	-
Width Mid-height	WMH	3,37E-07	0,3039	1,12E-07	0,0536	0,3074	-
Maximum Width	MW	3,29E-07	0,301	1,26E-07	0,0566	0,2894	-
Maximum Height	MH	1,04E-06	0,8235	2,94E-03	0,804	0,1158	-
Fruit Shape Index							
Fruit Shape Index external I	FSII	3,61E-11	0,6114	1,E-08	0,7793	0,1513	-
Blockiness							
Proximal Fruit Blockiness[1]	PFB	1,62E-08	0,01258	0,0001807	0,6499	0,09434	-
Distal Fruit Blockiness	DFB	0,01827	0,2313	0,0005386	0,1888	0,05399	-
Fruit Shape Triangle	FST	0,00399	0,8396	0,03264	0,4925	0,00105	2018-2019: 0,0109* 2018-2020: 0,4644 2019-2020: 0,0013**
Proximal Fruit End Shape							
Proximal Angle Macro[1]	PAMa	2,20E-16	0,03987	7,90E-13	0,2683	0,1103	-
Distal Fruit End Shape							
Distal Angle Macro[1]	DAMa	6,99E-09	0,02111	7,E-07	0,2887	8,58E-01	-
Internal Eccentricity							
Eccentricity	ECC	0,1203	0,2966	0,952	0,8933	2,25E-13	2019-2018: 3,1E-6*** 2020-2018: 0,0000*** 2020-2019: 0,0091**
FSI Internal	FSIINT	2,49E-08	0,2481	0,09396	0,2419	3,81E-05	2019-2018: 0,1041 2020-2018: 2,04E-5*** 2020-2019: 0,0309*
Homogeneity							
Ellipsoid[1]	E	9,51E-06	7,70E-06	0,1364	0,05074	5,15E-16	2019-2018: 0,9682 2020-2018: 0,0000*** 2020-2019: 0,0000*
Circular	C	0,1439	0,07349	0,2027	0,9082	3,66E-11	2019-2018: 0,9939 2020-2018: 0,0000*** 2020-2019: 0,0000***
Rectangular	R	0,002574	0,4645	0,0002854	0,4725	0,09791	-

[1]Measures adjusted in the number of genotypes and years evaluated, PFB with 90 genotypes, PAMa with 91 genotypes, E with 92 genotypes, and for DAMa was analyzed with subset 2018 and 2019 for this analysis. Derived data logarithmized. Derived from data logarithmized and reduce number genotypes. Derived from data not logarithmized and reduce number genotypes.

Derived from data logarithmized

Derived from data logarithmized and reduce number genotypes

Derived from data not logarithmized and reduce number genotypes

Supplementary Table 1.4. Heritability across years

Sort	Variable	Rep	Geno	Env	Year	Mean	Std	Min	Max	V.g	V.gxy	V.e	h2.s	h2.c	h2.p
1	**A**	1	94	1	3	35,15	6,70	25,26	56,92	36,43	0,00	25,88	**0,81**	0,81	0,82
2	**WMH**	1	94	1	3	7,35	0,68	5,49	9,84	0,37	0,24	0,27	**0,69**	0,81	0,81
3	**MW**	1	94	1	3	7,39	0,69	5,54	9,88	0,38	0,24	0,28	**0,69**	0,81	0,81
4	**MH**	1	94	1	3	6,08	0,67	4,87	8,29	0,37	0,22	0,27	**0,70**	0,82	0,82
5	**FSII**	1	94	1	3	0,82	0,07	0,70	1,09	0,00	0,00	0,00	**0,82**	0,92	0,93
6	**PFB**	1	94	1	3	0,75	0,02	0,71	0,80	0,00	0,00	0,00	**0,30**	0,75	0,75
7	**DFB**	1	94	1	3	0,69	0,02	0,63	0,73	0,00	0,00	0,00	**0,48**	0,83	0,83
8	**FST**	1	94	1	3	1,10	0,05	0,98	1,24	0,00	0,00	0,01	**0,38**	0,78	0,74
9	**PAMa**	1	94	1	3	261,55	22,40	106,04	297,06	270,86	306,88	431,71	**0,52**	0,70	0,72
10	**DAMa**	1	94	1	3	276,76	17,83	224,04	302,92	172,38	193,47	2757,86	**0,15**	0,58	0,57
11	**ECC**	1	94	1	3	0,49	0,05	0,38	0,62	0,00	0,00	0,01	**0,40**	0,76	0,77
12	**FSIINT**	1	94	1	3	0,51	0,08	0,35	0,76	0,00	0,00	0,01	**0,57**	0,86	0,87
13	**E**	1	94	1	3	0,12	0,01	0,09	0,15	0,00	0,00	0,00	**0,38**	0,68	0,69
14	**C**	1	94	1	3	0,18	0,03	0,12	0,24	0,00	0,00	0,00	**0,62**	0,84	0,85
15	**R**	1	94	1	3	0,56	0,01	0,52	0,58	0,00	0,00	0,00	**0,35**	0,78	0,78

Supplementary Table 1.5. Meteorological data over a three-year period.

Month	Average daily temperature + hour		
	2018	2019	2020
4	13,4	12,5	13,8
5	16,7	16	19,1
6	21,8	22,2	20,8
7	25,2	25,2	24,8
8	24,3	24,4	24,3
9	21,5	19,9	19,7
10	14,7	15,4	13,2
Mean	19,657143	19,3714286	19,3857143

Month	Maximum daily temperature + hour		
	2018	2019	2020
4	29,1	24,6	25,3
5	29,7	30	33,8
6	36,1	42,7	34,6
7	36,3	38,9	38,9
8	38,7	36,4	38,3
9	34,3	32	34,3
10	28,3	29,2	26,6
Mean	33,214286	33,4	33,1142857

Month	Minimum daily temperature + hour		
	2018	2019	2020
4	0,2	2	2,3
5	1,7	-0,2	8,5
6	11,1	6,4	7,9
7	13,5	12,1	10,9
8	12,4	10,5	9,1
9	7,3	8,2	5,9
10	-0,3	4,9	-0,4
Mean	6,5571429	6,27142857	6,31428571

Month	Accumulated daily precipitation		
	2018	2019	2020
4	78,6	34,1	70
5	60	46	65,7
6	12,8	8,7	77,7
7	23	59,6	4,1
8	33,7	16,1	17,2
9	19,3	7,3	4,2
10	98,8	78,7	13
Mean	46,6	35,7857143	35,9857143

Month	Reference evapotranspiration		
	2018	2019	2020
4	89,12	92,31	89
5	124,37	137,44	135,71
6	146,99	166,69	151,53
7	176,55	167,32	172,69
8	151,05	146,9	145,58
9	107,81	105,91	89,72
10	64,19	64,5	52,12
Mean	122,86857	125,867143	119,478571

CHAPTER 2

Genome Wide Association Studies for size and shape measures in apple fruit.

Dujak, Christian[1] and Aranzana, Maria José[1,2]

[1] Centre for Research in Agricultural Genomics (CRAG) CSIC-IRTA-UAB-UB, Campus UAB, 08193, Bellaterra, Barcelona, Spain.

[2] IRTA (Institut de Recerca i Tecnologia Agroalimentàries), 08140, Caldes de Montbui, Barcelona, Spain.

ABSTRACT

Genomic tools facilitate the efficient selection of genetic materials with interesting traits within a breeding program. In this work, two traits of interest for apple fruit quality, shape and size, were studied. Metric data of 11 fruit morphology parameters, obtained in 355 genotypes of the Apple REFPOP collection (an apple reference collection representative of the genetic variability of European cultivated apples) in three years of harvest, were used for genome-wide association analysis (GWAS). We used two models for the analysis, FarmCPU and BLINK. The analysis identified 59 SNPs associated with fruit size and shape traits (35 with FarmCPU and 45 with BLINK) responsible for 71 QTNs. The QTNs were distributed in all chromosomes but in chromosome 10 and 15. Thirty-four QTNs, identified by 27 SNPs, were related for size traits and thirty-seven QTNs, identified by 26 SNPs, were related to shape attributes. The definition of the haploblocks containing the most relevant SNPs served to propose candidate genes, among them the genes of the ovate family protein MdOFP17 and MdOFP4 which were in a 9.7-kb haploblock on chromosome 11. RNA-seq data revealed low or null expression of these genes in the oblong cultivar (SKO) and higher expression in the flat (GRA). Further studies will be required to validate the role of the most promising markers and/or genes in natural variation for their ultimate use in breeding.

Keywords: GWAS, apple fruit shape QTNs, apple fruit size QTNs, haploblock analysis, ovate family proteins (OFP).

INTRODUCTION

Domesticated apples belong to the diploid species *Malus x domestica* (Suckow) Borkh (Coart et al., 2006; Harrison & Harrison, 2011; Ordidge et al., 2018), with a haploid chromosome number x = 17 and a highly duplicated genome of 651 Mb on size (Daccord et al., 2017). The high genetic diversity of apple cultivars and their high heterozygosis (Peace et al., 2019; Jung et al., 2020) are behind a wide phenotypic diversity. Apple breeding programs have traditionally considered fruit quality and productivity within their key objectives, however the need of novel varieties adapted to the effects driven by the change of the climatic conditions (water scarcity, higher temperatures, incidence of novel or emerging diseases, among others) demands novel and more efficient breeding strategies. Such strategies include novel phenotyping methods as well as the use of molecular markers (molecular breeding) (Laurens et al., 2018).

In apple breeding, the use of different strategies based on molecular markers allowed efficient selection (Migicovsky et al., 2021; Fazio, 2021). Molecular breeding is being progressively adopted in commercial breeding programs, while previous scientific development is always required. For such development, scientists require materials as well as phenotypic data and genomic tools.

An important tool for such purpose in apple is the REFPOP, a European collection of apple cultivars made of 534 genotypes (accessions and progenies) representing the current European breeding germplasm. This collection was genotyped with high density SNP arrays and phenotypically evaluated over years. To study the environment effect on the genotypes, the collection was copied in six countries (Jung et al., 2020). Jung et al., (2022), using the REFPOP phenotypic and genotypic data, conducted genome-wide association studies (GWAS) and genomic selection (GS). The genotypic data consisted

on a genome-wide high-density dataset of 303,239 SNPs while phenotypic information referred to numerous traits, including flowering time, harvested date, productivity, and fruit traits such as color, russeting, bitter pit, and fruit size. This study allowed for the identification of relevant QTNs that need to be validated for their use in breeding.

Apart from the study mentioned above, over the last years, several works have aimed at the identification of DNA polymorphisms associated with apple traits. Chagné et al., (2019) put together a list of 128 single nucleotide polymorphisms (SNPs) for their validation in a panel of accessions, which included commercial varieties, advanced selections, and seedlings. Some of the SNPs were highly associated with the trait, so can be efficiently used for molecular breeding.

Most of the works have been addressed to identify markers associated to disease resistance genes and to fruit quality traits like color, acidity, firmness, or compounds related to flavor. Although several publications have focused on fruit size and shape, which are important fruit quality traits, the identification of genes or genomic regions regulating these traits in apples and the development of markers for marker assisted selection (MAS) is still a challenge.

Regarding the genetic inheritance and regulation of fruit shape, most of advances have been done in vegetable crops. For example, Mauxion et al., (2021) review the knowledge of the cellular and molecular mechanisms controlling fruit size in tomato, which is largely considered as an excellent model to study fruit growth and development, as well as fruit size. Several genes (*SUN*, *OVATE*, *FS8.1*) have been described as genes/QTL controlling the ovary and fruit elongation in tomato (Wu et al., 2015; Wang et al., 2019; Mauxion et al., 2021). In apple, some significant molecular markers (SSRs and SNPs) for fruit size and shape traits (diameter, length, and height of the fruit) have been located along the

whole apple genome (with exception of chromosome 6) by QTL mapping (Kenis et al., 2008; Devoghalaere et al., 2012; Chagné et al., 2014; Potts et al., 2014; Costa, 2015; Sun et al., 2015; Liu et al., 2016), however their efficiency for marker assisted selection is low. Also, only few shape QTLs, considered as the ration between width and heigh (i.e. fruit shape index, FSI), have been identified in segregating populations (Sun et al., 2012; Chang et al., 2014; Cao et al., 2015).

Therefore, to increase the knowledge on genomic regions, markers and genes involved in the inherited natural variation of fruit morphology, in this work we used the fruit measures obtained for a broad description of fruit shape and size in Chapter 1. In total 15 measurements were considered, four for size and 11 for shape. Values were obtained in three consecutive years in genotypes of the apple REFPOP copy maintained in Lleida. GWAS and RNA-seq data served to identify and propose candidate genes that will require further validation.

MATERIALS AND METHODS

Plant Material

The apple REFPOP was used to analyze the fruit shape and size measurements. This collection is described by Jung et al., (2020). We analyzed 355 genotypes consisting of 257 accessions and 98 seedlings derived from 31 families (Supplementary Data 2.1). We collected at least two replicates per genotype of the REFPOP copy grown in Lleida, Spain. The treatment conditions of the field for each year were: pruning season, thinning, iron chelate as fertilizer, not hormones application, and daily watering.

Phenotyping and data analyses

The dataset derived from previous analysis in **Chapter 1**. All but one of the measures (Proximal Fruit Blockiness) were consider in this work. The measurements were dimensional (size parameters) and dimensionless (measures for shape) and were acquired in fruits harvested in three (from 2018 to 2020). Measures evaluated seven fruit aspects: size, fruit shape ratio, blockiness, homogeneity, distal fruit end shape, and internal eccentricity, as provided by the Tomato Analyzer software Version 3 by Gonzalo et al., (2009). In addition, we used in the analysis the categorical classification of CAT-own of Chapter 1, which sorts the fruits in three groups: oblate or flat, spheroid or round, and oblong. The dataset is provided in **Supplementary Data 2.2**.

Phenotypic raw data was obtained for at least three apples per clone (two clones per genotype) and year. The mean values for each genotype were used for the analyses, having 134 genotypes in 2018, 274 genotypes in 2019, and 339 genotypes in 2020. In addition, in Chapter 1 the variance among the years evaluated was analyzed. For the genotypes with more than one year of data, the mean values were calculated for each measure, obtaining a final dataset of 355 genotypes (referred as mean across-years dataset). Spearman's correlation for all datasets, the distribution of the data and the density plots and heatmaps were calculated and plotted with the *ggplot2* package (Wickham, 2016) in R Core Team (2022) program.

Genotyping data

Genotypic data is reported in Jung et al., (2022), a total of 303,239 biallelic SNPs derived from markers imputation of two sets: the Affymetrix Axiom® Apple 480K SNP genotyping array (Bianco, et al., 2016) and the Illumina Infinium® 20K SNP genotyping

array (Bianco et al., 2014), corresponding to the accessions and seedlings genotypes. This previous analysis was based on the apple reference genome doubled haploid GDDH13 v1.1 (Daccord et al., 2017) to define their chromosomal positions.

Genome-wide association studies

Two methods were applied for this association analysis. The Fixed and random model Circulating Probability Unification method (Liu et al., 2016) (FarmCPU) combines the Mixed linear model with the Fixed Effect Model (FEM), removes the cofounding from Kinship, and reduces false negatives. The Random Effect Model (REM) selects associated markers by maximum likelihood method avoiding the over-fitting. The Bayesian-information and Linkage-disequilibrium Iteratively Nested Keyway method (Huang et al., 2018) (BLINK), replaces REM with FEM using the Bayesian information criteria (BIC) based on the linkage disequilibrium, in this way the method generates fewer false positives and high statistical power. Both methods were implemented in the R package GAPIT 3.0 (Wang and Zhang, 2021). GWAS was analyzed with different genomic matrices, including the same number of markers, 303,239 SNPs, and four datasets n=134 genotypes in 2018, n=274 genotypes in 2019, n=339 genotypes in 2020, and n=355 genotypes with mean-across year values. For both methods, we used three principal components and filtered out for the minor allele frequency (MAF) < 0.05. The Bonferroni correction was used to identify the p-value markers with significance threshold a = α/m with α = 0.05, m=number of markers being (–$log10(p$-value) > 6.75). The p-values were represented in multiple Manhattan-plot and QQ-plot using the threshold (Yin, 2018). Significant QTNs for all datasets and methods were graphically represented along each chromosome (Wickham, 2016). Additionally, the output GAPIT file provided the

phenotype variance explained by SNP (PVE) and the minor allele frequency (MAF). In addition, the coefficient of determination between phenotypes-genotypes was calculated (coding the alleles numerically, 1 and 2 correspond to homozygous alleles and 3 to heterozygous alleles). The allelic frequency of each significant SNP was calculated with its corresponding association (phenotype), represented in boxplot (Wickham, 2016).

Haploblocks-LD

Linkage disequilibrium (LD) was analyzed, creating Haploblock using Haploview software (Barrett et al., 2005), derived from previous filtering using PLINK (Purcell et al., 2007) at 100kb (+/-) from the position of significant SNPs, based on the position of the GDDH13 v1.1 genome (Daccord et al., 2017). The criteria for haploblocks were Hardy Weinberg p-value cutoff 0.01, minimum genotype 75%, maximum Mendel error one, and minimum minor allele frequency 0.05. To determine the blocks, we used the criteria of Gabriel et al., (2002), where the minimum confidence interval for strong LD (D') at the top 0.95 and at the bottom 0.2 (indicating the LD level from 0.2 to 1). The allelic frequency of each haplotype in the population and the connections from one block to another was also calculated with Haploview software.

Candidate genes annotation

Genes in the haploblocks or in a 200kb region flanking the candidate SNPs 100kb both sides were annotated based on the HFTH1 genome. Also, the haploblock regions initially delineated by their position in the GDDH13v1.1 genome were subsequently aligned to the HFTH1 genome by BLAST+ from GDR database (Jung et al., 2019). Posteriorly we used databases such as Gene Ontology (GO) terms (Ashburner et al., 2000), InterPro

(IPR) (Hunter et al., 2009), Kyoto Encyclopedia of genes and genomes (KEGG orthologs and pathways) (Kanehisa et al., 2016), non-redundant proteins sequences from NCBI (RefSeq) (Pruitt et al., 2007), Arabidopsis thaliana orthologs from the Arabidopsis Information Resource (TAIR) (Lamesch et al., 2012) and computer-annotated protein sequence database for the translation of coding sequences (UniProtKB/TrEMBL) (The UniProt Consortium, 2019).

RNA extraction and cDNA preparation

Sampling collected three different apple fruit shapes and sizes: flat-large, round-medium/small, and oblong-medium at 13 days after anthesis into the fruit development with their three biological replicates. All samples were processed with liquid nitrogen and stored at -80ºC. The samples corresponded to 'Grand'mere' (GRA), 'Kansas Queen' (KAN), and 'Skovfoged '(SKO) varieties from the apple REFPOP (Jung et al., 2020) collection located in Lleida, Spain. Total RNA was extracted with Maxwell® RSC simplyRNA tissue kit, using Maxwell® RSC instrument and was purify twice with Turbo® DNase. Checked the quality and quantify with Bioanalyzer® system, and was sent to Novogene (London, England).

For validation of RNA-seq data, RNA samples were converted to cDNA using the PrimeScript RT Reagent Takara kit with the first step components consisting of Oligo(dt) 20 nt (50 uM), total RNA, H2O RNase-free for 5 min at 70ºC. The second step RT reaction was performed, including 5X PrimeScript Buffer, PrimeScript RT Enzyme, RNase out, dNTPs (100 uM), plus adding the first step and H2O RNase-free incubating at 50ºC for 60 min and inactivation at 70ºC for 15 min. RT-PCRs were visualized by 1.5% agarose gel and 1X TAE for verification of the integrity and subsequent use.

Bioinformatic analyses

Sequencing libraries filtering, mapping to reference genome and quality control

The mRNA sequencing libraries were filtered by sequence quality (reads with a Phred score < 30 were removed), and Illumina sequencing adaptors remained were trimmed by using Trim-Galore (Krueger et al., 2012) version 0.6.1. Burrows-Wheeler Aligner (Li and Durbin, 2010). High-quality RNA sequencing libraries were also mapped to HFTH1 by using HISAT (Kim et al., 2015) version 2.1.0. with default settings parameters.

Statistical data of mapped and unmapped reads in Binary Alignment Map (BAM) files were analyzed using SAMStat (Lassmann et al., 2011) version 1.5.1. The index used to determine the quality of alignment and assembly was the Mapping Quality Score (Li et al., 2009)(MAPQ), which quantifies the probability of a misplaced read. Assembled libraries in BAM format were filtered by MAPQ score, based on the quality of the alignment. Multiple aligned reads (reads with MAPQ< 30 or not properly aligned according to the mapper) were filtered, and statistic reports were obtained with Samtools (Li et al., 2009) version 1.9. Besides, Samtools was used to transform, index, and sort the files generated by the mappers according to the protocol's needs.

Quality control reports were obtained before and after filtering and mapping with FastQC (Andrew, 2010) version 0.11.5 to ensure high-quality standards for downstream analyses. All the quality reports were summarized in an Html file by using MultiQC (Ewels et al., 2016) version 1.9.

The gene quantification and count matrix were constructed with featureCounts (Liao et al., 2014) setting the parameters to paired-end sequencing, avoiding chimeric count fragments (those fragments that have their two ends aligned to different chromosomes), specifying as feature exon feature type for reading counting and

annotated as transcript and allowing overlapping features for the differential use of exon during alternative splicing. The results obtained were normalized to Transcript Per Million (TPM).

The batch effect was checked with the *sva* R package version 3.12 (Leek et al., 2012). In addition, preliminary exploratory analysis and visualization of the samples from the dataset were performed. For count matrix normalization, a regularized-logarithm transformation (*rlog*) was applied, recommended to stabilize the variance across the mean for negative binomial data with a dispersion-mean trend and a low number of samples ($n < 30$).

Validation RNA-seq and expressed genes (TPM)

For RNA-seq validation we designed primers based on HF43536 mRNA sequences, Fw 5'-AGGGCAGCTAAGGATTTGGA-3' and Rv 5'-TGTGTGTGCCATGTCAAACCAG-3'. The qPCR was performed using the LightCycler 480 System Roche, qPCR components were 5x MasterMix SYBR Green, primers Fw and Rv (each 10uM), H20 nuclease-free and cDNA adjusted to dilution 1:40. The qPCR efficiency was 2.0 and the R-squared between log2(TPM) and Ct was 0.86. Once validated, the expressed genes annotated according to haploblock were analyzed, first by testing the distribution of the data (Shapiro test) and then, depending on whether data was normally or non-normally distributed we used Anova-one way or Kruskal-Wallis, respectively. We applied a confidence level of $p<0.05$ and determined differences between genotypes the Tukey HSD test using the normalized count matrix in TPM.

PhenoGram

PhenoGram (Wolfe et al., 2013), based on chromosomal ideograms sharing the genomic information, were constructed using significant SNPs or any molecular markers obtained from association analysis and were visualized in a physical map; in that case, the markers published were aligned using BLAST-NCBI (Madden et al., 2003) with referent genome double haploid GDHH13 v1.1 (Daccord et al., 2017). All QTNs obtained in this study were included, and molecular markers published about apple fruit shape and size measures.

RESULTS

In this work we used metric data for fruit morphology traits obtained with the Tomato Analyzer software images scanned of 12,692 apple sections. The apples corresponded to 355 genotypes collected over three years (134 in 2018, 274 in 2019 and 339 in 2020; 93 common between years) (**Supplementary Data 2.1**). The traits evaluated, which referred to size and shape attributes, are broadly described in **Chapter 1**. As shown in **Supplementary Figure 2.1 and Data 2.2**, for most variables, data showed similar distribution for, at least, two of the years. In addition to the morphometric data obtained with the Tomato Analyzer software we included in the analysis the description of the fruits in three classes: oblate (flat), spheroid (round), and oblong (elliptic), named as CAT-own classification as described in **Chapter 1.**

Correlation values of attributes obtained between years, the mean across-years values, and the year attributes with the mean across-years values are shown in **Figure 2.1 and Supplementary Figure 2.2.**

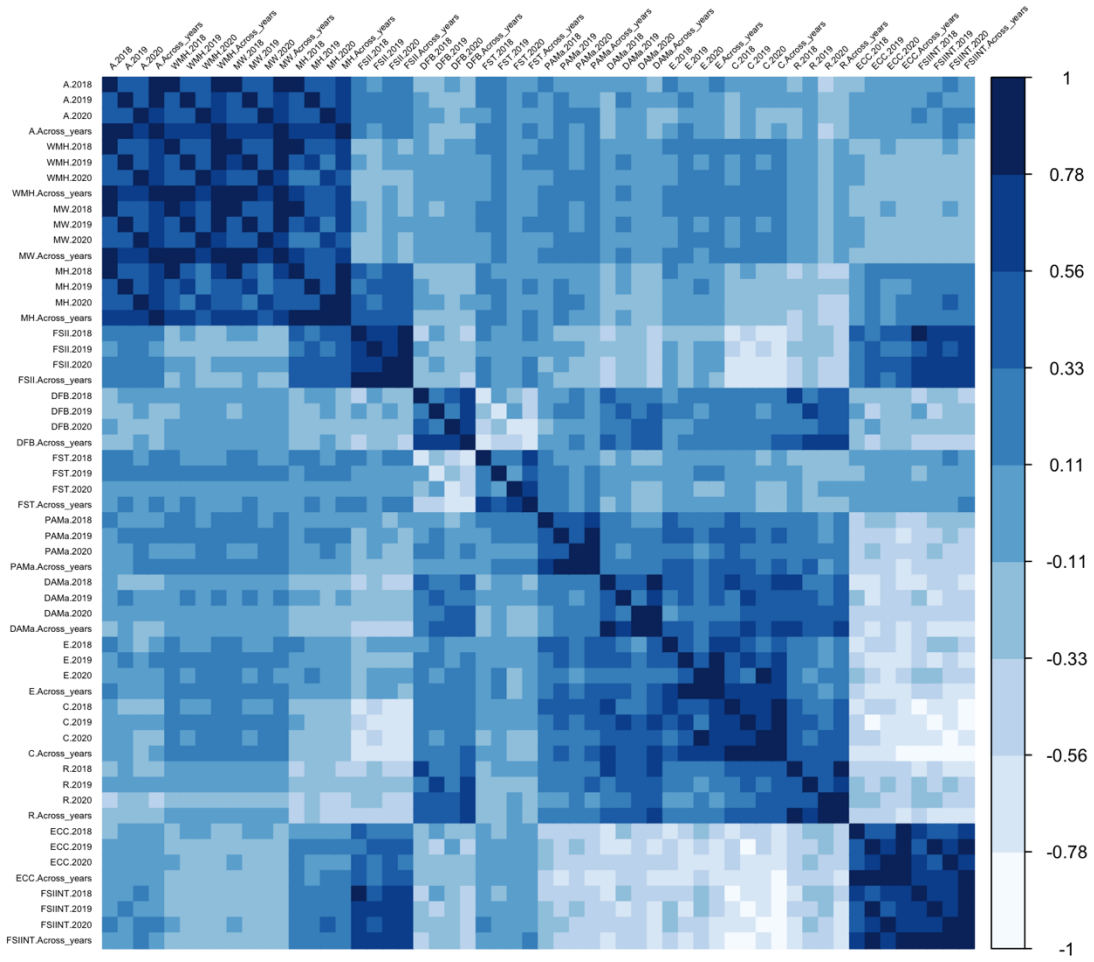


Figure 2.1. Spearman correlation between size and shape phenotypic datasets. Data was obtained in 2018, 2019, 2020 and a fourth dataset was generated with the mean across years values. The acronymous for the phenotypic attributes correspond to Area (A), Width Mid-height (WMH), Maximum Width (MW), Maximum Height (MH) Fruit shape index external I (FSII), Distal fruit blockiness (DFB), Fruit shape triangle (FST), Proximal angle macro (PAMa), Distal angle macro (DAMa), Ellipsoid (E), Circular (C), Rectangular (R), Eccentricity (ECC) and Fruit shape internal (FSIINT). See details in **Supplementary Figure 2.2**.

Most of the attributes showed from moderate to high correlation between years. When considering the correlations between year and the mean values for a given attribute, the lowest value was found for the FST observations in 2019 (r=0.51) while the highest correlation was observed for FSIII in 2020 (r=0.91) (**Supplementary Figure 2.2**).

Genome wide association studies

Genome-wide association studies (GWAS) were conducted for all traits using the per year as well as the mean across-years values using two models (FarmCPU and BLINK)

(**Figure 2.2**). The results are shown in the Manhattan plot and QQ-plot in the **Supplementary Figures 2.3 and 2.4**, respectively. The analysis found SNPs with association values over the Bonferroni threshold (-log10(p) =6.751) for all traits but for the fruit shape triangle (FST), for the distal angle macro (DAMa), ellipsoid (E) and for the eccentricity (ECC). Considering the two GWAS models, the three years of data and the mean values, we identified 59 SNPs associated (35 with FarmCPU and 45 with BLINK) responsible for 71 QTNs (**Figure 2.2**) i.e., different QTNs (22 in total) were identified by the same SNP (10). Most of the QTNs (39) were found when using the mean values. Five QTNs were identified simultaneously with two datasets (in all cases they were detected with the 2020 and with the mean values datasets) and nine QTNs were identified by the two models in either one of the year's assessments (six QTNs) or when using the means (three QTNs).

In total, seven SNPs were simultaneously associated with more than one attribute, being one of the SNPs associated with three (AX-115482211 in chromosome 2, with A, MW and MWH). The 71 QTNs were distributed along all but the 10 and 15 chromosomes, ranging from two to 13 QTNs per chromosome. While some QTNs were scattered along the chromosome, others were in clusters.

This was the case, for example, of 10 QTNs distributed in a 248 kb region at the top of the chromosome 11 for FSII, FSIINT and CAT-own. Downstream that chromosome we found a region of 554 kb with three QTNs for C, MW and WMH, and finally at the bottom of the chromosome three QTNs of FSII in a 524 kb region (**Supplementary Table 2.1**).

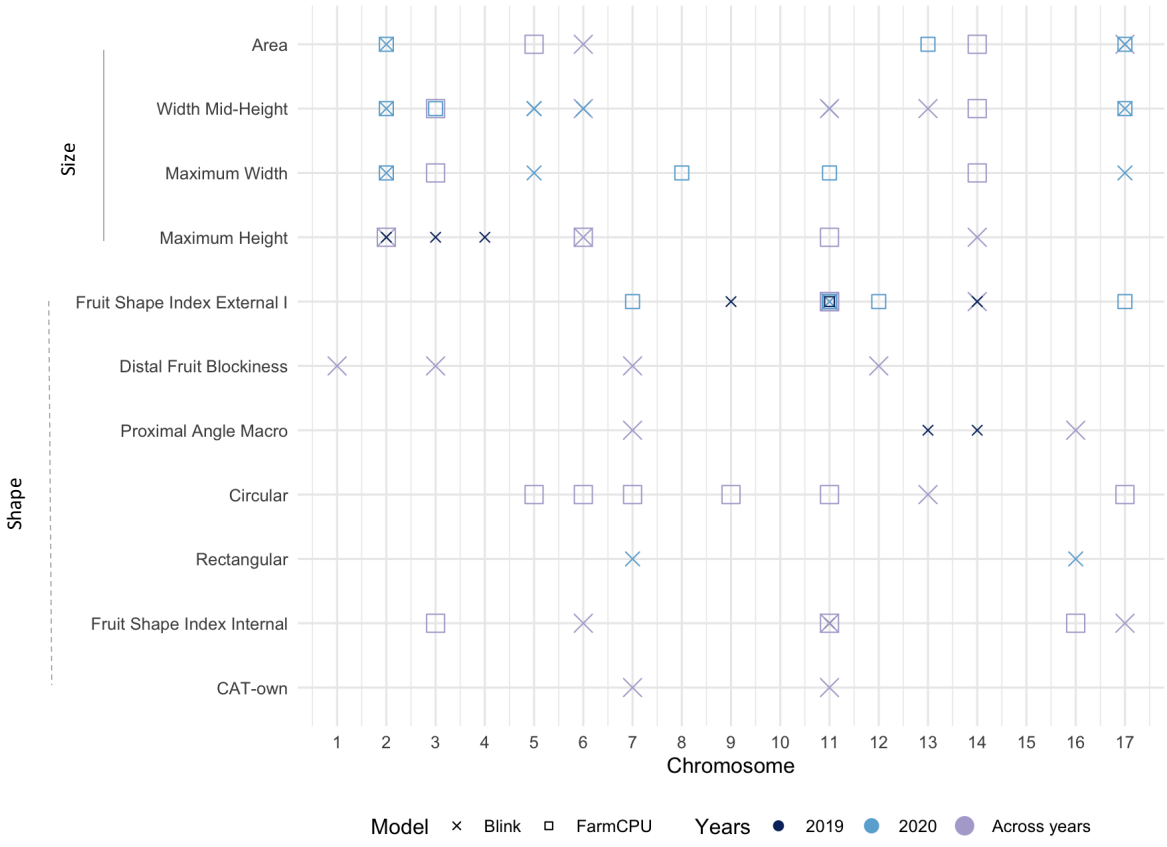


Figure 2.2. Summary of GWAS results for years evaluated with different models for size and shape measures. QTNs obtained with Blink (x) and FarmCPU (□) models in the 2019 (●), 2020 (●) and mean values (●) datasets are represented. Chromosomes on the x-axis and measurements on the y-axis.

Quantitative Trait Nucleotides for size-related traits

In total we found 34 QTNs for size-related traits: 12 for the width mid-height (WMH), eight for the maximum height (MH), seven for the maximum width (MW) and seven for the area (A) (**Figure 2.2 and Supplementary Table 2.1**). Most of the QTNs were found for the 2020 and/or for the mean data. These QTNs involved 27 SNPs, identifying five of them more than one QTN: the SNPs AX-115482211 (chromosome 2) and AX-115481999 (chromosome 3) identified three QTNs each; and the SNPs AX-115378078 (chromosome 6), AX-115295642 (chromosome 14), and AX-115312607 (chromosome 17) identified two QNTs each. Four of these SNPs were simultaneously significant for MW and WMH QTNs (at chromosomes 2, 3, 14, 17); the one at chromosome 2 (AX-115482211) was also

significant for the A trait (**Figure 2.2 and Supplementary Table 2.1**). The QTNs for the width-mid height were found in eight chromosomes.

Quantitative Trait Nucleotides for shape-related traits

We found 37 QTNs for shape-related traits distributed in twelve chromosomes: twelve QTNs for the fruit shape index external I (FSII), seven for the circular measure (C), six for the fruit shape index internal (FSIINT), four for each measure (proximal angle macro and distal fruit blockiness), and two QTNs for CAT-own and rectangular values. Three SNPs (AX-115335214 and AX-105213957 32 Mb apart in chromosome 11, and AX-115336086 in chromosome 14) were responsible for six FSII QTNs for either 2020 or the mean values datasets. One SNP was associated to two QTNs (CAT-own and FSII) and one SNP to C and FSIINT. In the chromosome 11, we identified 9 QTNs for C, FSII, FSIINT and CAT-own distributed along the chromosome, although 4 QTNs (one for FSIINT, one for CAT-own and two for FSII) were in a region of 248kb (**Supplementary Table 2.1**).

Twenty-three of the QTNs were found when using the mean values, and 14 QTNs with the 2019 and 2020 datasets. Four out of the thirty-two SNPs were found simultaneously with the 2020 and the mean values datasets (**Supplementary Table 2.1**).

Phenotypic variation and haploblocks

The phenotypic variation explained (PVE) by each SNP ranged from 0.03% to 12.51%, with a mean of 3.74 % (**Supplementary Table 2.2**). For each SNP and QTN, we used violin plots for the graphic visualization of genotype-phenotype distributions. The distributions of 14 depicting phenotypic differences between the three genotypic classes (homozygous for the reference allele, heterozygous, and homozygous for the alternative allele) are shown in **Figure 2.3**.

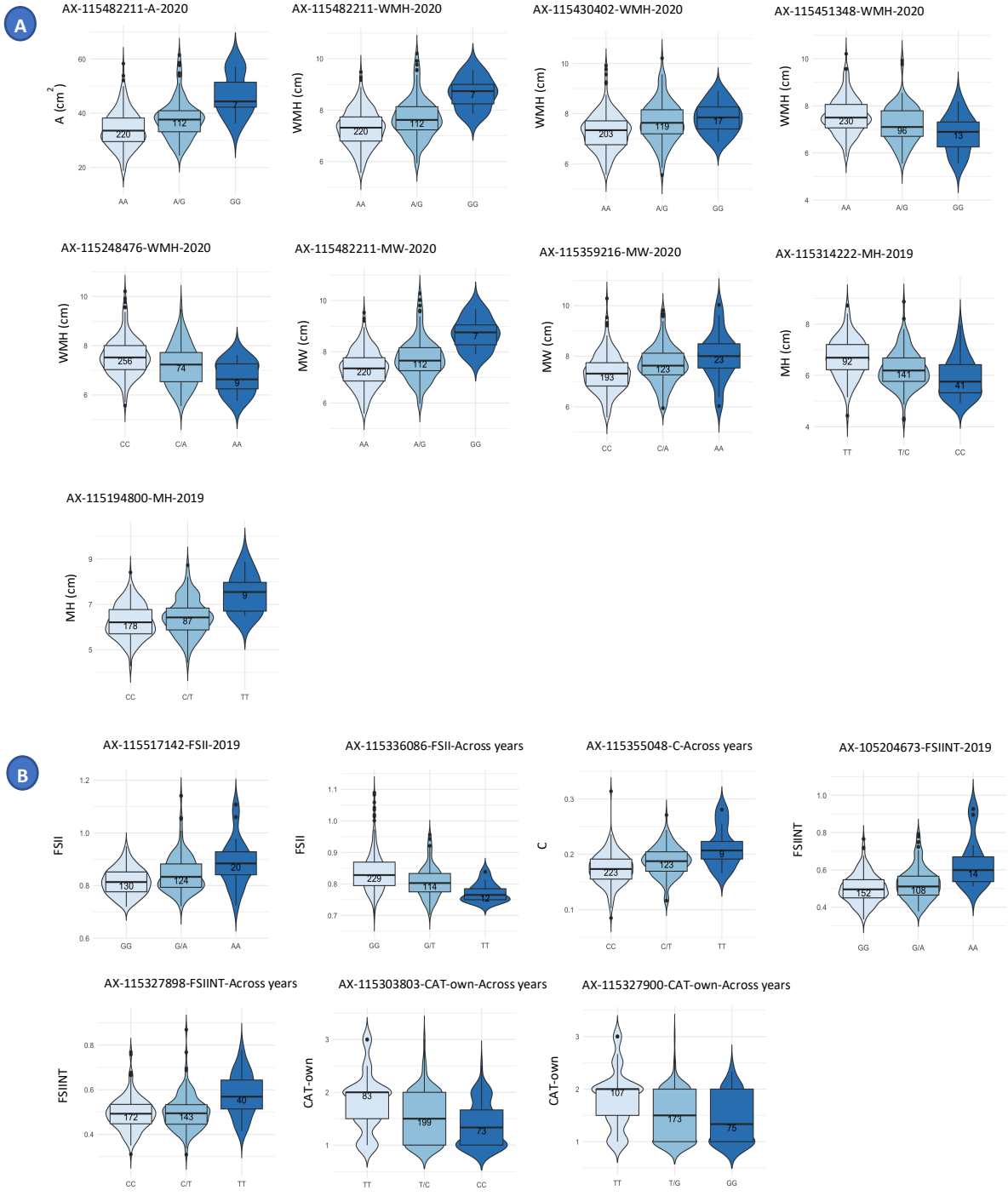


Figure 2.3. Violin plots showing the frequency distribution of size (A) and shape (B) phenotypic values across genotypes. Each violin plot corresponds to the QTN-Trait and their frequencies between them. The first allele (on the left) represents the homozygous genotype for the reference allele (GDDH13v1.1), in the middle the heterozygous genotype and on the right the homozygous genotype for the alternative allele.

One of the SNPs, the AX-115482211 SNP, was simultaneously associated with three fruit size measures (Area, Width-mid height and Maximum width). Individuals with the alternative allele G (allele frequency of 19%) produced fruits with larger area, maximum width, and width-mid height (**Figure 2.3A** and **Supplementary Table 2.3**). The effect was observed in heterozygous as well as in homozygous individuals. The phenotype distribution for additional size and shape associated markers are represented in **Figure 2.3A and 2.3B, and Supplementary Table 2.3**.

For the 10 most outstanding SNPs we constructed haploblocks. As two of the SNPs were linked, we obtained nine resulting haploblocks. Those were in six chromosomes (2, 4, 6, 7, 11, and 13) and had an average size of 31.5 kb, ranging from 1.1 to 111 kb. Thirteen QTNs occurred in these haploblocks (**Supplementary Table 2.4**).

Numerous QTNs for size and shape attributes were identified in a window of 1.9 Mb along the chromosome 11. Eleven of them were in haploblocks (linked or co-segregating) (**Figure 2.4**). In **Figure 2.5** we summarize the results on three relevant shape-related QTNs and their causal SNPs: AX-115327898 (C/T alleles) and AX-115327900 (T/G alleles), both 5kb apart at the top of chromosome 11 and associated to FSIINT and CAT-own attributes, respectively; and AX-115355048 in chromosome 13 and associated to the circular measure provided by the Tomato Analyzer software, which describes how well the fruit section depicts a circle (**Figure 2.5B**).

The cultivars with the allele T in AX-115327898 in homozygosis (11.3%) showed higher FSIINT values (i.e. were preferably oblong), trend that was not observed in the heterozygous individuals.

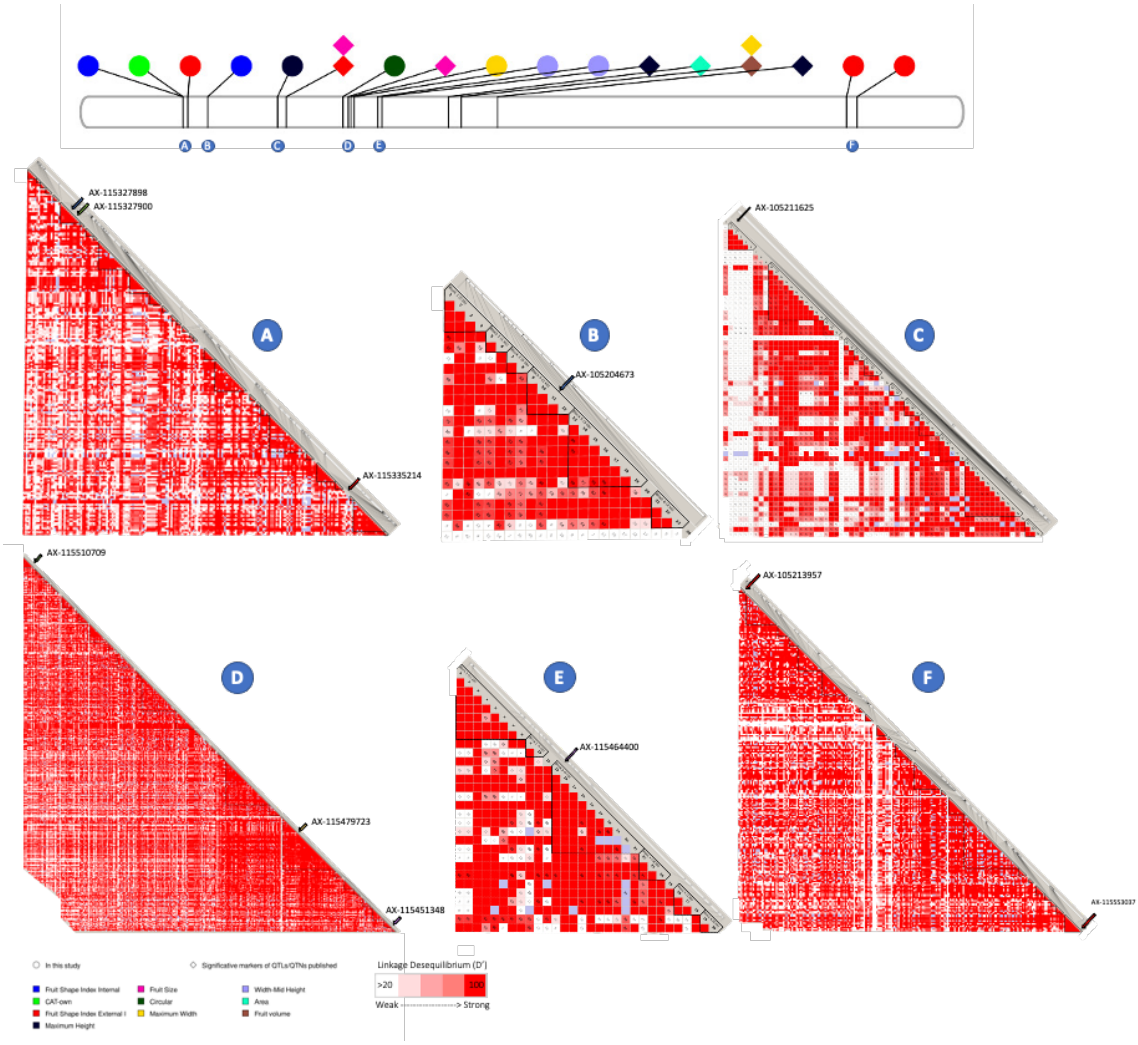


Figure 2.4. Linkage disequilibrium on Chromosome 11 GDDH13v1.1. On top figure corresponds to chromosome 11 from GDDH13v1.1 position of markers published and in this study. Symbols: circle are QTNs found in this study and kite are markers published, the colors represent to QTN-Trait and each circle and letters correspond to the haploblocks. Haploblock using GDDH13v1.1 position: Haploblock A: 4.661.534 to 4.958.286 (199 markers), Haploblock B: 5.930.883 to 5.948.789 pb (24 markers), Haploblock C: 9.046.670 to 9.556.662 pb (63 markers), Haploblock D: 12.638.696 to 13.198.304 pb (434 markers), Haploblock E: 14.355.224 to 14.402.233 pb (30 markers), Haploblock F: 37.648.782 to 38.174.120 bp (241 markers). Symbols: white to red represent at level of linkage disequilibrium (D') expressed in percentage, >20 (weak) and red is 100 of D'.

By contrary, individuals with CC and CT genotypes at this site (89%) produced flat and circular fruits (accessions such as 'Grand'mere' and 'Kansas Queen'). The second SNP (AX-115327900) was associated with the Cat-own categorical classification. The allele G (with a frequency of 44%) was associated with flat and circular apples ('Grand'mere' and 'Kansas Queen') with effect observed in the homozygous (20%) as well as in the heterozygous (48%) individuals, while those with the TT genotype (30%) tended to show oblong shapes (such as the 'Skovfoged' variety) (**Figure 2.5C**). These two SNPs were in complete LD (D'=1) and occurred in a haploblock 9,7 kb long, which had seven haplotypes with an average frequency of 0.14, ranging from 0.02 to 0.428 (**Figure 2.5B**). The haploblock around the AX-115355048 SNP in chromosome 13 (with C/T alleles) was significant for the circular (C) attribute, with 10 haplotypes 18,7 kb long with frequencies ranging from 0.378 to 0.011. The allele T was observed with higher frequency when the inner area of the fruit had higher C values, and therefore, deviated more from the circularity (see **Figure 2.5C**) towards flat shape, for FSIINT and Circular values are inversely correlated (see **Figure 2.1**). The homozygous TT genotypes (with average Circular = 0.22 and FSIINT=0.77 values), despite being at low frequency in the collection (3%), and heterozygous CT such as in 'Kansas Queen' (average Circular=0.20 and FSIINT=0.80 values), had higher C. By contrary, homozygous CC showed lower Circular value, and therefore higher FSIINT (average Circular = 0.19 and FSIINT=0.83).

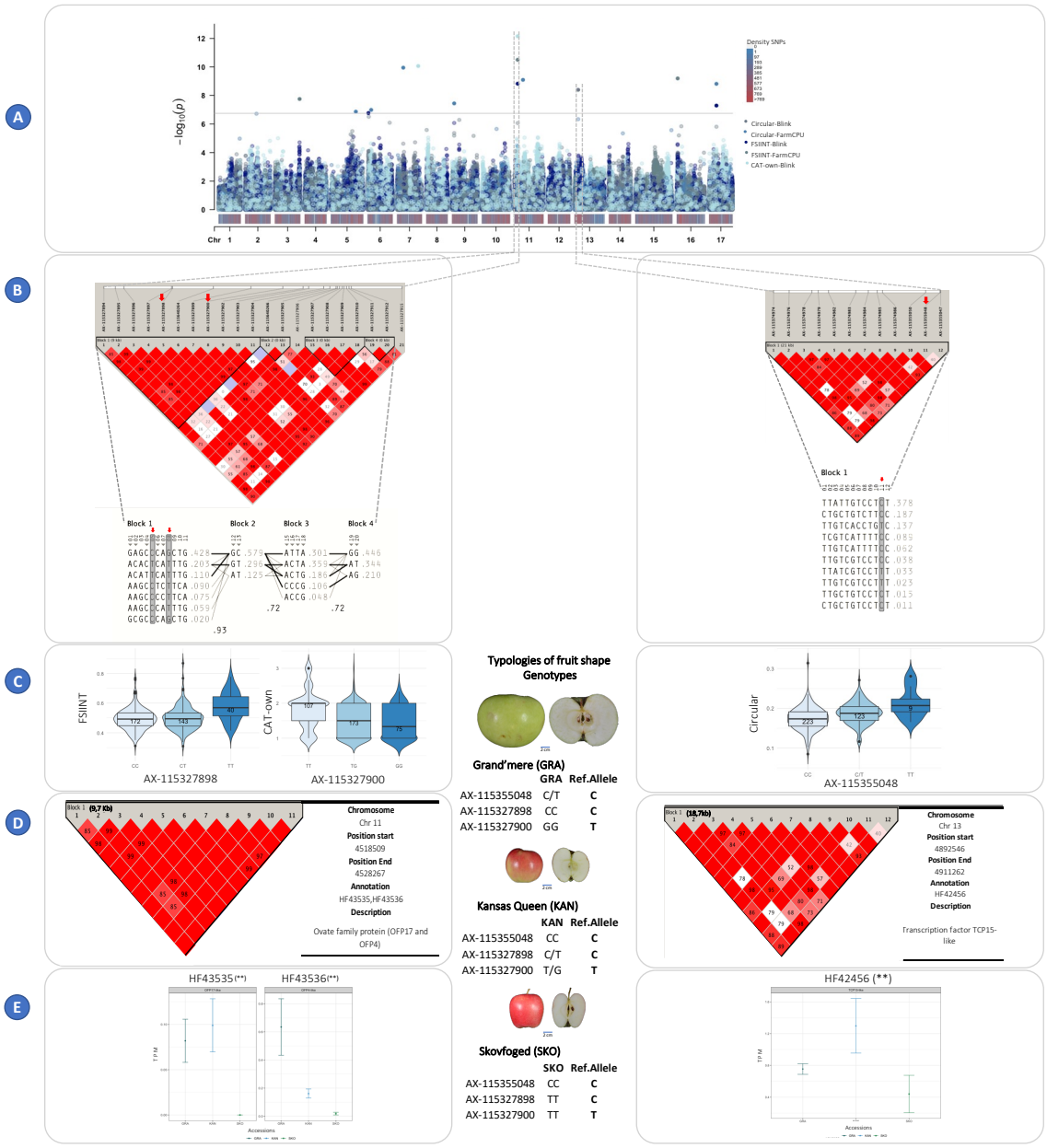


Figure 2.5. Summary of the global analysis of the QTNs for FSIINT, CAT-own and Circular with the population across years (355 genotypes). Above is the **A, GWAS results:** it includes the multiple Manhattan plots of the traits FSIINT, CAT-own and Circular with the two models used (Blink and FarmCPU). The colors represent trait and model. Density plots indicate the density of SNPs on each chromosome, as indicated in the legend of the graph. **B, Haploblocks & Haplotypes**: show the linkage disequilibrium (D') based on GDDH13v1, the criteria were based on Gabriel et al. (2002) strong LD (0,9) and weak (0,2), the colors indicate the level of D' (white= weak and red= strong). Below are the haplotypes of each block and the allelic frequency of the block. **C, Frequency Genotype-Phenotype:** allele frequency of three traits and their alleles. The middle of the figure shows three representative genotypes of apple shape: 'Grand'mere' (flat),' Kansas Queen' (round) and 'Skovfoged' (oblong) and their genotypes according to SNPs. **D,**

Gene annotation: haploblock represents the candidate region for gene annotation, using the annotations of the HFTH1v1 genome. **E, RNA-seq:** Transcripts per million (TPM) for the three genotypes mentioned above. HF43535, HF43536 and HF42456 genes analyzed in the three genotypes, the samples were collected from fruits at 13 Days After anthesis (DAA) stage.

Gene annotation

For each of the 59 associated SNPs we searched for the genes annotated in a 200 kb region (100 Kb up- and down-stream the SNP position) in HFTH genome. This annotation identified 873 genes were annotated, 371 genes corresponded to size QTNs and 502 to shape QTNs. The 53% of the annotations contained molecular description, according to GO databases. Fifty-one genes were described to have protein-binding molecular function, 40 genes related to biological processes (as for example regulation of transcription, DNA repair, phosphorylation, and transmembrane transport among others), as well as genes involved in cell division, growth, cell modification, response to hormones (gibberellin, auxin, and ethylene). According to TAIR database, some genes were related to fruit development and growth, nine genes were related to auxin response (HF06172, HF40493, HF29276, HF02793, HF08237, HF41541, HF02644, HF02646, HF12008), four genes to ethylene response (HF14170, HF14173, HF16534, HF11991), three genes to gibberellin-regulated (HF41950, HF38795, HF08230), two genes already described for fruit shape such as ovate family protein (HF43535, HF43536) (**Supplementary Table 2.5**).

We also explored the gene annotation in the 9 haploblocks previously defined based on the linkage disequilibrium, identifying a total of 30 genes according to the TAIR database. Notably, among these we found the ovate family genes 17 and 4 (HF43535, HF43536), the *TCP15*-like transcription factor involved in plant regulation (HF42456),

and several proteins of the kinase superfamily (**Figure 2.5D and Supplementary Table 2.4**).

RNA-seq gene expression (TPM)

Whole RNA sequence data of three genotypes, one oblate ('Grand'mere'), one round ('Kansas Queen') and one oblong ('Skovfoged') obtained from fruits at 13 days after anthesis were analyzed to evaluate the expression in fruit of the 30 genes annotated in the haploblocks (**Supplementary Table 2.5**). Twenty-three of them were transcriptionally expressed in fruits of the three genotypes. The differential expression levels between genotypes of the 23 transcripts were analyzed by Kruskal-Wallis and Anova test, using the count matrix normalized in transcripts per million (TPM). In total, six genes were differentially expressed between genotypes: the genes HF43535 and HF43536 (ovate family protein 17 and 4) annotated in the haploblocks of the SNPs AX-115327898 and AX-115327900 in chromosome 11; the genes HF10079 and HF10080 (Patched family and protein kinase proteins, respectively) in the haploblocks of the SNPs AX-115513701 and AX-115448691 SNPs in chromosome 6; the gene HF15994 (unknown function protein) in the haploblock of the SNP AX-115194800 in chromosome 4; the gene HF42456 (transcription factor TCP15-like) in the haploblock of the SNP AX-115355048 in chromosome 13.

According to this annotation, three genes had a function in organ regulation and development: *Ovate family protein genes* HF43535 *and* HF43536 and *Transcription factor TCP15-like gene HF42456.*

Ovate family protein genes HF43535 *and* HF43536

The differential expression of both genes was highly significant at a confidence level of $p<0.01$. We also applied Tukey-HSD test to check for differences between pairs of

genotypes. For the HF43535 gene (ovate family protein 17), significant differences in expression were observed between the oblate variety 'Grand'mere' and the oblong 'Skovfoged' (GRAvsSKO) and between 'Kansas Queen' (round) and 'Skovfoged' (KANvsSKO), while not between the oblate and round varieties (GRAvsKAN). This gene was expressed at a lower level in the oblong variety 'Skovfoged' (**Figure 2.5E and Supplementary Table 2.6**).

The HF43536 gene (ovate family protein 4) was differentially expressed in the pairs GRAvsKAN (oblate and round) and GRAvsSKO (oblate and oblong), with higher RNA levels in the oblate genotype.

As a mean to validate the RNA-seq data, the gene expression of this last gene (HF43536) was assessed by RT-qPCR, obtaining a r-squared of 0.8591 between Ct and log2(TPM) values (**Supplementary Figure 2.5**).

Transcription factor TCP15-like gene HF42456

This transcription is involved in the regulation of plant development. Differences in gene expression levels were found between the oblate and oblong fruits (GRAvsSKO) and between round and oblong fruits (KANvsSKO), with the SKO genotype showing lower gene expression (**Figure 2.5E and Supplementary Table 2.6**).

DISCUSSION

Fruit shape, and in particular the shape of the apple, is relevant both in the description and varietal characterization (as can be seen from its inclusion in the ECPGR and UPOV guides for the characterization and registration of varieties, respectively), as well as in aspects related to its commercialization. While in breeding programs the vague concept "nice shape" is included among the selection criteria, international marketing directives

regulate aspects that are not much more specific, such as those referred to as "shape defects" normally associated with poor fruit development (OECD, 2021).

In addition to the shape, the size of the fruit has great commercial relevance, not only at the productive but also at the market level, for larger fruits are marketed as "Extra" and "Class I" categories with higher market value.

Although visual and "easy to evaluate" criteria as the FSII for fruit classification is useful for the above-mentioned purposes, more objective data is necessary to perform genome studies. Here we used the data and measures obtained and described in **Chapter 1** to search for genomic regions controlling apple fruit shape and size attributes. Fruit size and shape data was obtained in thousands of images acquired in fruits of a total of 355 genotypes in three consecutive harvesting campaigns (93 genotypes were common in the three years of assessments). It is broadly accepted that climatic and management factors affect fruit shape and size, although some studies show low differences in the FSII ratio between years, only observed under high divergences in the air and soil temperature in spring, for it may cause differences in the fruit seed number, the main factor determining fruit shape (Tromp, 1990). Spring temperatures were only moderately milder in spring 2020, compared to 2018 and 2019, so we shouldn't expect extreme divergences. Indeed, in chapter 1 we found from moderate to high heritability values (H2), in particular for the shape-related traits. In addition, the correlations between the values taken each year support this fact.

Some of the attributes measured were highly positively or negatively correlated, as shown in **Figure 2.1 and Supplementary Figure 2.2**.

Although to date the crucial genes regulating apple fruit size and shape have not been identified, several published studies have aimed at the identification of relevant QTLs. If

we put together the markers identified in this work with markers from published in other studies (Kenis et al., 2008; Chang et al., 2014; Potts et al., 2014; Cao et al., 2015; Jung et al., 2022) and construct a PhenoGram with 110 molecular markers (SNPs and SSRs; 76 markers for size traits and 37 for shape) in **Figure 2.6** we can see all markers were aligned accordingly to the physical map of the genome version GDDH13v1.1. Chromosome 11 contained the highest number of markers (19), followed by chromosomes 2, 5 and 17 (13 each one), the other markers were distributed in the remaining chromosomes. In addition, many of the markers identified in this study were located close to other published markers (between 0.4 to 2 Mb in distance) to (**Supplementary Figure 2.6 and Table 2.7**).

Mapping for size traits

We found several SNPs associated with size-related traits. In chromosome 3, Potts et al., (2014) found two QTLs (for fruit circumference and height) in a segregating population, with an explained phenotypic variation of 45%. At 3.8 Mb distance we found two QTNs (for MW and WMH) in the vicinity of the gene HF40493, an auxin response factor 4 homologous to *AtARF4* (Want et al., 2020), which regulates both female and male gametophyte development in *Arabidopsis* (Liu et al., 2018) .

In chromosome 4, Chang et al., (2014) reported a QTL for fruit height close to the SSR Hi23g08, although with a low LOD (2.01) and contribution to the total variance (6.2). In this chromosome we only found a significant SNP for maximum heigh at a 18 Mb distance (AX-115194800) with the allele alternative to the reference associated to genotypes with elongated fruits. The PVE for this SNP and trait was low (1.57%).

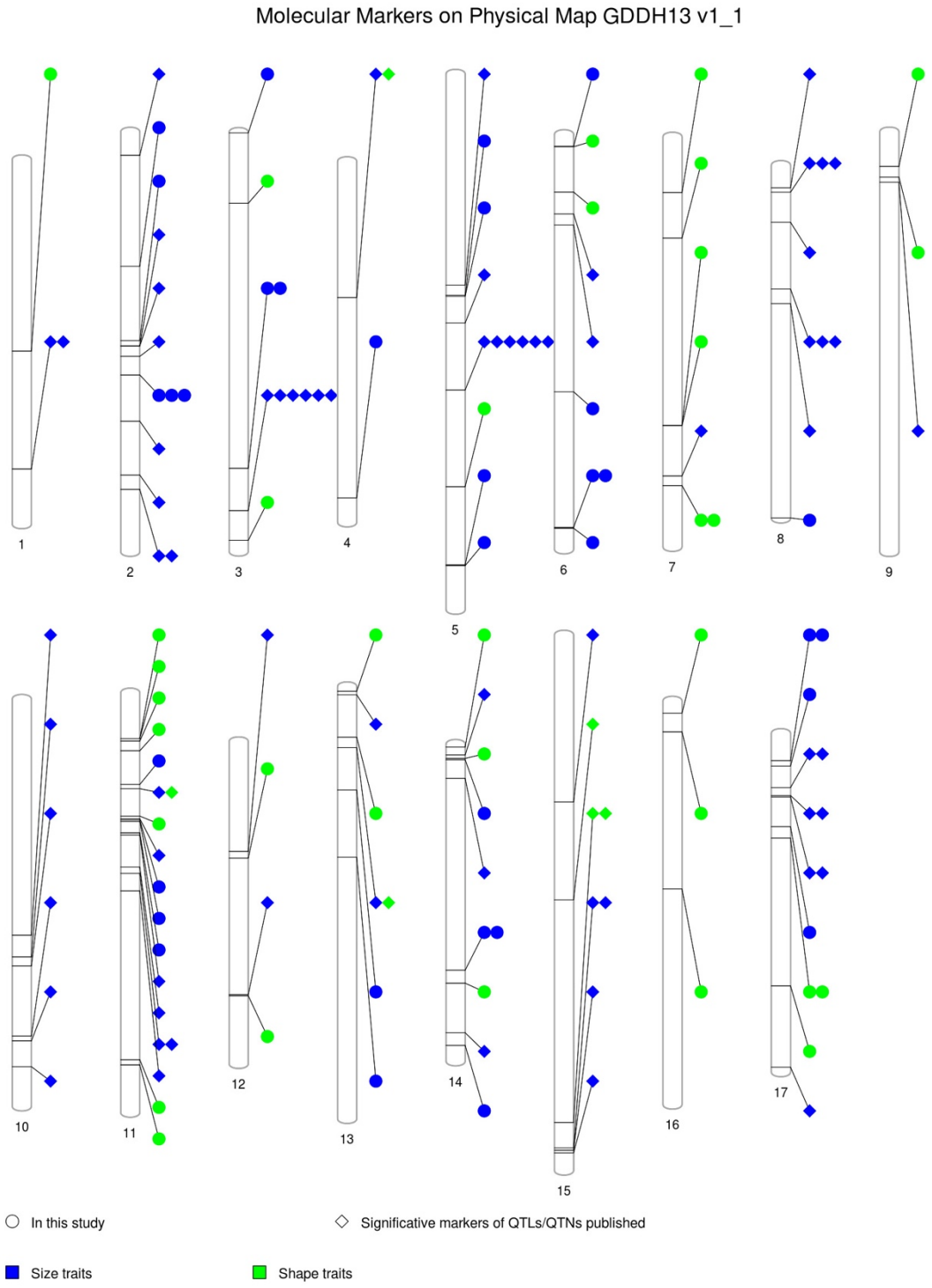


Figure 2.6. PhenoGram of the molecular markers on Physical map according to the GDDH13v1.1 apple genome sequence. Significant markers mapped for apple fruit measures, including markers published in QTLs/QTNs analysis and the QTNs found in this study. Symbols: circle, correspond to "in this study" and kite, "significative markers of QTLs/QTNs published". Color blue for Size traits and green for Shape traits. See details in **Supplementary Figure 2.6**.

On chromosome 5, we found eight QTNs. Two of them (one for A and one for WMH) were at a very close distance (82 kb a part). The two responsible SNPs (AX-115248476 and AX-115435503) were in LD (R-squared mean was 0.38) and added up to 10.6% of PVE. These QTNs were at about 15.7 Mb from a QTL for fruit maximum heigh (Potts et a, 2014) detected in one of the two years of evaluation only, with a LOD of 3.94 and 21.5% of the variance explained.

Two other QTNs for fruit width were at the top of this chromosome with the two SNPs (AX-115638603 and AX-115436710) 102 kb apart and at less than 1Mb from a QTL for the same attribute identified by Chang et al., (2014) with a LOD of 2.9 explaining 9.2% of the variance, and 2.3 Mb apart from a QTL also for fruit width identified by Kenis et al., (2008) with LOD3.5 and 12.4% of the variance explained. The SNPs explained together 6.57% of the variance. Similarly, Potts et al., (2014) identified a width QTL 8.4Mb downstream of the SNPs. All together confirms the existence of a width-responsible region in chromosome 5. According to the TAIR database description, the annotated genes in these regions are responsible for growth regulation such as Transcriptional factor B3 family protein/auxin-responsive factor AUX/IAA-related (HF12008), ethylene responsive element binding factor 1 (HF11991), Gibberellin-regulated family protein (HF08230) and Auxin-responsive GH3 family protein (HF08237). Based on the results of the GWAS annotations, genes involved in fruit development and growth were identified. Such as hormones, that play an important role in fruit growth and are controlled by multiple genes (Kumar et al., 2014), for example, in melon, two overlapping QTLs, one for fruit diameter and one for fruit weight were detected on chromosome 11, identifying the gene *MELO3C025758* (auxin response factor) as one of the candidate genes for these traits (Lian et al., 2021). In tomato, auxin and gibberellin

hormones regulate the transition from flower stage to fruit set (Jong et al., 2009). Endogenous auxin concentration is one of the factors controlling fruit size in apple (Bu et al., 2020). Devoghalaere et al., (2012) suggest a potential role in fruit size of the Auxine Responsible Factor (*ARF106*) gene, contained in a QTL for fruit weight on chromosome 15. In addition, these authors found other QTLs on chromosomes 5, 6, 8, 11, 12 and 17 for the same trait. Similarly, three QTNs (A and WMH) associated with size traits, separated by 84 kb at position 35 Mb on chromosome 6 were identified, as well as on the same chromosome have been identified with the same traits by Jung et al., (2022) and by Kenis et al., (2008) with MH.

Yao et al., (2015) validated the negative effect of the overexpression of the miRNA172 in apple fruit size. This miRNA was in the confidence interval of a fruit size-related QTL in chromosome 11. Chang et al., (2014) also found QTLs for apple size in the same region. Here, we identified one QTN for fruit height at 216 kb distance from the SNP AX-115464400 in Jung et al., (2022). In addition, QTLs/QTNs for fruit size attributes have also been reported along chromosomes 2, 8, 13, 14, and 17 (Kenis et al., 2008; Devoghalaere et al., 2012; Chang et al., 2014; Jung et al., 2022) as in this study.

Mapping for shape traits

In tomato, several genes for the control of fruit shape have been identified, such as *OVATE*, S*UN*, *FAS* (fasciated) and *LC* (Locule number). The *SUN* and *OVATE* genes control shape elongation, while *FAS* and *LC* control locule number and flat shape (Tanksley, 2004;Brewer et al., 2007; Rodriguez et al., 2011). In pepper the fw2.1 locus co-localizes with the *OVATE* gene and is associated with smaller fruit (Zygier et al., 2005), also in cucumber, three QTLs for fruit shape have been identified, one of them (fruit shape index 2.1) with a phenotypic variation greater than 50% (Gao et al., 2020). In melon, the

QTL fsqs8.1 is associated with the round shape of the fruit, at whose locus the gene *CmOFP13* (ovate family protein) is annotated (Martinez-Martinez et al., 2022). In peach, a 1.7 Mb downstream inversion of the gene encoding the ovate *PpOFP1* is responsible for the flat shape (Zhou et al., 2021).

In this study, several QTNs for apple fruit shape have been identified by association with six dimensionless measurements, among them the association with FSIINT and CAT-own located within a haploblock (9.7 kb) on chromosome 11 at position 4.6 Mb. Two ovate family protein genes (*MdOFP17* and *MdOFP4*) are annotated in this region where two SNP markers have been found highly associated the flat, round, or oblong shapes.

The QTN for the Circular (C) measurement was identified within a 18.7kb haploblock on chromosome 13, the alternative allele of the SNP (AX-115355048) was associated with genotypes bearing fruits with tendency to the oblong shape. In addition, the *TCP15*-like trancription factor gene, the only annotation of the haploblock, is involved in the regulation of plant development and the stimulation of biosynthesis of hormones such as brassinosteroid, jasmonic acid, and flavonoids (Li, 2015).

As seen, the fruit shape index (FSII) measure has been used in successive works, as well as in this one, since it is the measure with higher weight in the definition of fruit shape (See **Chapter 1**). Within these regions associated with shape traits in apple, candidate genes such as the ovate protein family are identified.

In total, 11 QTNs were identified on chromosome 11 (for FSII and FSIINT). Cao et al., (2015) have reported a QTL for the same measure (FSII), at a distance of 5 Mb from SNPs (AX-115327898 and AX-115327900) associated with FSII and another measure (CAT-own) that are highly correlated. Chang et al., (2014) also detected several QTLs for fruit

shape index (FSI), one of those QTLs in LG11 contributed to a phenotypic variance between 10.3 - 13.7% in a segregating population.

Genes expressed by RNA-seq data

Ovate family protein

The expression of the *OFP17* and *OFP4* genes, located in the same haploblock, was analyzed in the three genotypes, whose phenotypes represent three types of apple shape, showing that the expression of both genes in the 'Skovfoged' genotype (with oblong fruits) is lower or null than in the other genotypes. In contrast, in the 'Grand'mere' genotype (large and flat fruit) the expression of the *OFP4*-like gene is higher than in the other genotypes.

The ovate family proteins are genes involved in the regulation of plant development in different organs described in several species such as Arabidopsis, tomato, melon and peach. They are transcriptional repressor genes, but they also play an important role in the regulation of cell division in tomato fruit development, or in response to hormone changes (Wang et al., 2016; Snouffer et al., 2020). In Arabidopsis, tomato and rice, the over-expression of OFPs causes the cotyledon, fruit and seed to be flattened or, if there is a mutation in these genes, the organs are elongated (Liu et al., 2002; Wang et al.,2007; Schmitz et al., 2015; Yang et al., 2016;2018).

In apples, the diversity of OFP genes (26) distributed in 13 chromosomes has been studied (Li et al., 2019), but their role in apple fruit shape has not been described yet.

To date, few studies have been carried out to know which candidate regions or genes are responsible for fruit size and shape, but in this study, we present QTNs and candidate genes for a better understanding and contribution of molecular markers for breeding.

CONCLUSION

In this study, numerous genomic regions associated with 11 measurements related to quantitative traits of apple fruit size and shape have been identified. The REFPOP collection from Europe provided a diversity of genotypes revealing the different associations between markers and traits. Although size and shape traits are polygenic, genes such as OFPs and several hormones described in fruit development have been identified. This study also provides markers that could be used in a breeding program for assisted selection.

REFERENCES

1. Ashburner, M., Ball, C. A., Blake, J. A., Botstein, D., Butler, H., Cherry, J. M., ... & Sherlock, G. (2000). Gene ontology: tool for the unification of biology. *Nature genetics*, *25*(1), 25-29.
2. Barrett, J. C., Fry, B., Maller, J. D. M. J., & Daly, M. J. (2005). Haploview: analysis and visualization of LD and haplotype maps. *Bioinformatics*, *21*(2), 263-265.
3. Bianco, L., Cestaro, A., Linsmith, G., Muranty, H., Denancé, C., Théron, A., ... & Troggio, M. (2016). Development and validation of the Axiom® Apple480K SNP genotyping array. *The Plant Journal*, *86*(1), 62-74.
4. Bianco, L., Cestaro, A., Sargent, D. J., Banchi, E., Derdak, S., Di Guardo, M., ... & Troggio, M. (2014). Development and validation of a 20K single nucleotide polymorphism (SNP) whole genome genotyping array for apple (Malus× domestica Borkh). *PloS one*, *9*(10), e110377.
5. Brewer, M. T., Moyseenko, J. B., Monforte, A. J., & van der Knaap, E. (2007). Morphological variation in tomato: a comprehensive study of quantitative trait loci controlling fruit shape and development. *Journal of experimental botany*, *58*(6), 1339-1349.
6. Bu, H., Yu, W., Yuan, H., Yue, P., Wei, Y., & Wang, A. (2020). Endogenous auxin content contributes to larger size of apple fruit. *Frontiers in plant science*, *11*, 592540.

7. Cao, K., Chang, Y., Sun, R., Shen, F., Wu, T., Wang, Y., ... & Han, Z. (2015). Candidate gene prediction via quantitative trait locus analysis of fruit shape index traits in apple. *Euphytica*, *206*(2), 381-391.
8. Chagné, D., Dayatilake, D., Diack, R., Oliver, M., Ireland, H., Watson, A., ... & Tustin, S. (2014). Genetic and environmental control of fruit maturation, dry matter and firmness in apple (Malus× domestica Borkh.). *Horticulture Research*, *1*.
9. Chagné, D., Vanderzande, S., Kirk, C., Profitt, N., Weskett, R., Gardiner, S. E., ... & Bassil, N. V. (2019). Validation of SNP markers for fruit quality and disease resistance loci in apple (Malus× domestica Borkh.) using the OpenArray® platform. *Horticulture research*, *6*.
10. Chang, Y., Sun, R., Sun, H., Zhao, Y., Han, Y., Chen, D., ... & Han, Z. (2014). Mapping of quantitative trait loci corroborates independent genetic control of apple size and shape. *Scientia Horticulturae*, *174*, 126-132.
11. Coart, E. L. S., Van Glabeke, S., De Loose, M., Larsen, A. S., & ROLDÁN-RUIZ, I. (2006). Chloroplast diversity in the genus Malus: new insights into the relationship between the European wild apple (Malus sylvestris (L.) Mill.) and the domesticated apple (Malus domestica Borkh.). *Molecular Ecology*, *15*(8), 2171-2182.
12. Costa, F. (2015). MetaQTL analysis provides a compendium of genomic loci controlling fruit quality traits in apple. *Tree genetics & genomes*, *11*(1), 1-11.
13. Daccord, N., Celton, J. M., Linsmith, G., Becker, C., Choisne, N., Schijlen, E., ... & Bucher, E. (2017). High-quality de novo assembly of the apple genome and methylome dynamics of early fruit development. *Nature genetics*, *49*(7), 1099-1106.
14. De Jong, M., Mariani, C., & Vriezen, W. H. (2009). The role of auxin and gibberellin in tomato fruit set. *Journal of experimental botany*, *60*(5), 1523-1532.
15. Devoghalaere, F., Doucen, T., Guitton, B., Keeling, J., Payne, W., Ling, T. J., ... & David, K. M. (2012). A genomics approach to understanding the role of auxin in apple (Malus x domestica) fruit size control. *BMC Plant Biology*, *12*(1), 1-15.

16. Ewels, P., Magnusson, M., Lundin, S., & Käller, M. (2016). MultiQC: summarize analysis results for multiple tools and samples in a single report. *Bioinformatics*, *32*(19), 3047-3048.
17. Fazio, G. (2021). Genetics, breeding, and genomics of apple rootstocks. In *The apple genome* (pp. 105-130). Springer, Cham.
18. Gabriel, S. B., Schaffner, S. F., Nguyen, H., Moore, J. M., Roy, J., Blumenstiel, B., ... & Altshuler, D. (2002). The structure of haplotype blocks in the human genome. *science*, *296*(5576), 2225-2229.
19. Gao, Z., Zhang, H., Cao, C., Han, J., Li, H., & Ren, Z. (2020). QTL mapping for cucumber fruit size and shape with populations from long and round fruited inbred lines. *Horticultural Plant Journal*, *6*(3), 132-144.
20. Gonzalo, M. J., Brewer, M. T., Anderson, C., Sullivan, D., Gray, S., & van der Knaap, E. (2009). Tomato fruit shape analysis using morphometric and morphology attributes implemented in Tomato Analyzer software program. *Journal of the American Society for Horticultural Science*, *134*(1), 77-87.
21. Harrison, N., & Harrison, R. J. (2011). On the evolutionary history of the domesticated apple. Nature genetics, 43(11), 1043-1044.
22. Heberle, H., Meirelles, G. V., da Silva, F. R., Telles, G. P., & Minghim, R. (2015). InteractiVenn: a web-based tool for the analysis of sets through Venn diagrams. *BMC bioinformatics*, *16*(1), 1-7.
23. Huang, M., Liu, X., Zhou, Y., Summers, R. M., & Zhang, Z. (2019). BLINK: a package for the next level of genome-wide association studies with both individuals and markers in the millions. *Gigascience*, *8*(2), giy154.
24. Hunter, S., Apweiler, R., Attwood, T. K., Bairoch, A., Bateman, A., Binns, D., ... & Yeats, C. (2009). InterPro: the integrative protein signature database. *Nucleic acids research*, *37*(suppl_1), D211-D215.
25. Inglis, P. W., Pappas, M. D. C. R., Resende, L. V., & Grattapaglia, D. (2018). Fast and inexpensive protocols for consistent extraction of high quality DNA and RNA from challenging plant and fungal samples for high-throughput SNP genotyping and sequencing applications. *PloS one*, *13*(10), e0206085.

26. Jung, M., Keller, B., Roth, M., Aranzana, M. J., Auwerkerken, A., Guerra, W., ... & Patocchi, A. (2022). Genetic architecture and genomic predictive ability of apple quantitative traits across environments. *Horticulture research*, *9*.
27. Jung, M., Roth, M., Aranzana, M. J., Auwerkerken, A., Bink, M., Denancé, C., ... & Muranty, H. (2020). The apple REFPOP—a reference population for genomics-assisted breeding in apple. *Horticulture research*, *7*.
28. Jung, S., Ficklin, S. P., Lee, T., Cheng, C. H., Blenda, A., Zheng, P., ... & Main, D. (2014). The genome database for Rosaceae (GDR): year 10 update. *Nucleic acids research*, *42*(D1), D1237-D1244.
29. Kanehisa, M., Sato, Y., Kawashima, M., Furumichi, M., & Tanabe, M. (2016). KEGG as a reference resource for gene and protein annotation. *Nucleic acids research*, *44*(D1), D457-D462.
30. Kenis, K., Keulemans, J., & Davey, M. W. (2008). Identification and stability of QTLs for fruit quality traits in apple. *Tree Genetics & Genomes*, *4*(4), 647-661.
31. Kim, D., Langmead, B., & Salzberg, S. L. (2015). HISAT: a fast spliced aligner with low memory requirements. *Nature methods*, *12*(4), 357-360.
32. Krueger, F., & Andrews, S. R. (2011). Bismark: a flexible aligner and methylation caller for Bisulfite-Seq applications. *bioinformatics*, *27*(11), 1571-1572.
33. Krueger, F., & Andrews, S. R. (2012). Quality control, trimming and alignment of Bisulfite-Seq data (Prot 57). *Department of Medicine, Hematology and Oncology, Domagkstr*, *3*(48149), 1-13.
34. Kumar, R., Khurana, A., & Sharma, A. K. (2013). Role of plant hormones and their interplay in development and ripening of fleshy fruits. *Journal of experimental botany*, *65*(16), 4561-4575.
35. Lamesch, P., Berardini, T. Z., Li, D., Swarbreck, D., Wilks, C., Sasidharan, R., ... & Huala, E. (2012). The Arabidopsis Information Resource (TAIR): improved gene annotation and new tools. Nucleic acids research, 40(D1), D1202-D1210.
36. Langmead, B., & Salzberg, S. L. (2012). Fast gapped-read alignment with Bowtie 2. *Nature methods*, *9*(4), 357-359.
37. Lassmann, T., Hayashizaki, Y., & Daub, C. O. (2011). SAMStat: monitoring biases in next generation sequencing data. *Bioinformatics*, *27*(1), 130-131.

38. Laurens, F., Aranzana, M. J., Arus, P., Bassi, D., Bink, M., Bonany, J., ... & Van de Weg, E. (2018). An integrated approach for increasing breeding efficiency in apple and peach in Europe. *Horticulture research*, *5*.
39. Leek, J. T., Johnson, W. E., Parker, H. S., Jaffe, A. E., & Storey, J. D. (2012). The sva package for removing batch effects and other unwanted variation in high-throughput experiments. *Bioinformatics*, *28*(6), 882-883.
40. Li, H., Dong, Q., Zhao, Q., & Ran, K. (2019). Genome-wide identification, expression profiling, and protein-protein interaction properties of ovate family proteins in apple. *Tree Genetics & Genomes*, *15*(3), 1-11.
41. Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., ... & Durbin, R. (2009). The sequence alignment/map format and SAMtools. *Bioinformatics*, *25*(16), 2078-2079.
42. Li, S. (2015). The Arabidopsis thaliana TCP transcription factors: a broadening horizon beyond development. *Plant signaling & behavior*, *10*(7), e1044192.
43. Lian, Q., Fu, Q., Xu, Y., Hu, Z., Zheng, J., Zhang, A., ... & Wang, H. (2021). QTLs and candidate genes analyses for fruit size under domestication and differentiation in melon (Cucumis melo L.) based on high resolution maps. *BMC plant biology*, *21*(1), 1-13.
44. Liao, Y., Smyth, G. K., & Shi, W. (2014). featureCounts: an efficient general purpose program for assigning sequence reads to genomic features. *Bioinformatics*, *30*(7), 923-930.
45. Liu, J., Van Eck, J., Cong, B., & Tanksley, S. D. (2002). A new class of regulatory genes underlying the cause of pear-shaped tomato fruit. *Proceedings of the National Academy of Sciences*, *99*(20), 13302-13306.
46. Liu, X., Huang, M., Fan, B., Buckler, E. S., & Zhang, Z. (2016). Iterative usage of fixed and random effect models for powerful and efficient genome-wide association studies. *PLoS genetics*, *12*(2), e1005767.
47. Liu, Z., Bao, D., Liu, D., Zhang, Y., Ashraf, M. A., & Chen, X. (2016). Construction of a genetic linkage map and QTL analysis of fruit-related traits in an F1 Red Fuji x Hongrou apple hybrid. *Open Life Sciences*, *11*(1), 487-497.

48. Liu, Z., Miao, L., Huo, R., Song, X., Johnson, C., Kong, L., ... & Yu, X. (2018). ARF2–ARF4 and ARF5 are essential for female and male gametophyte development in Arabidopsis. *Plant and Cell Physiology*, *59*(1), 179-189.
49. Love, M. I., Huber, W., & Anders, S. (2014). Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome biology*, *15*(12), 1-21.
50. Madden, T. (2003). The BLAST sequence analysis tool. *The NCBI handbook*.
51. Martínez-Martínez, C., Gonzalo, M. J., Sipowicz, P., Campos, M., Martínez-Fernández, I., Leida, C., ... & Monforte, A. J. (2022). A cryptic variation in a member of the Ovate Family Proteins is underlying the melon fruit shape QTL fsqs8. 1. *Theoretical and Applied Genetics*, *135*(3), 785-801.
52. Mauxion, J. P., Chevalier, C., & Gonzalez, N. (2021). Complex cellular and molecular events determining fruit size. *Trends in Plant Science*, *26*(10), 1023-1038.
53. Migicovsky, Z., Gardner, K. M., Richards, C., Thomas Chao, C., Schwaninger, H. R., Fazio, G., ...& Myles, S. (2021). Genomic consequences of apple improvement. *Horticulture research*, *8*.
54. OECD. (2021). Organisation for Economic Co-operation and Development. Available at: https://www.oecd.org/
55. Ordidge, M., Kirdwichai, P., Baksh, M. F., Venison, E. P., Gibbings, J. G., & Dunwell, J. M. (2018). Genetic analysis of a major international collection of cultivated apple varieties reveals previously unknown historic heteroploid and inbred relationships. *PLoS One*, *13*(9), e0202405.
56. Park, Y., & Wu, H. (2016). Differential methylation analysis for BS-seq data under general experimental design. *Bioinformatics*, *32*(10), 1446-1453.
57. Peace, C. P., Bianco, L., Troggio, M., Van de Weg, E., Howard, N. P., Cornille, A., ... & Vanderzande, S. (2019). Apple whole genome sequences: recent advances and new prospects. *Horticulture Research*, *6*.
58. Potts, S. M., Khan, M. A., Han, Y., Kushad, M. M., & Korban, S. S. (2014). Identification of quantitative trait loci (QTLs) for fruit quality traits in apple. *Plant molecular biology reporter*, *32*(1), 109-116.

59. Pruitt, K. D., Tatusova, T., & Maglott, D. R. (2007). NCBI reference sequences (RefSeq): a curated non-redundant sequence database of genomes, transcripts and proteins. *Nucleic acids research*, *35*(suppl_1), D61-D65.
60. Purcell, S., Neale, B., Todd-Brown, K., Thomas, L., Ferreira, M. A., Bender, D., ... & Sham, P. C. (2007). PLINK: a tool set for whole-genome association and population-based linkage analyses. *The American journal of human genetics*, *81*(3), 559-575.
61. Rodríguez, G. R., Muños, S., Anderson, C., Sim, S. C., Michel, A., Causse, M., ... & van Der Knaap, E. (2011). Distribution of SUN, OVATE, LC, and FAS in the tomato germplasm and the relationship to fruit shape diversity. *Plant physiology*, *156*(1), 275-285.
62. Schmidt, P., Hartung, J., Rath, J., & Piepho, H. P. (2019). Estimating broad-sense heritability with unbalanced data from agricultural cultivar trials. *Crop Science*, *59*(2), 525-536.
63. Schmitz, A. J., Begcy, K., Sarath, G., & Walia, H. (2015). Rice Ovate Family Protein 2 (OFP2) alters hormonal homeostasis and vasculature development. *Plant Science*, *241*, 177-188.
64. Snouffer, A., Kraus, C., & van der Knaap, E. (2020). The shape of things to come: ovate family proteins regulate plant organ shape. *Current opinion in plant biology*, *53*, 98-105.
65. Sun, H. H., Zhao, Y. B., Li, C. M., Chen, D. M., Wang, Y., Zhang, X. Z., & Han, Z. H. (2012). Identification of markers linked to major gene loci involved in determination of fruit shape index of apples (Malus domestica). *Euphytica*, *185*(2), 185-193.
66. Sun, R., Chang, Y., Yang, F., Wang, Y., Li, H., Zhao, Y., ... & Han, Z. (2015). A dense SNP genetic map constructed using restriction site-associated DNA sequencing enables detection of QTLs controlling apple fruit quality. *BMC genomics*, *16*(1), 1-15.
67. Tanksley, S. D. (2004). The genetic, developmental, and molecular bases of fruit size and shape variation in tomato. *The plant cell*, *16*(suppl_1), S181-S189.
68. Team, R. C. (2020). *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing.

69. Tromp, J. (1990). Fruit shape in apple under various controlled environment conditions. *Scientia horticulturae*, *43*(1-2), 109-115.

70. UniProt Consortium. (2019). UniProt: a worldwide hub of protein knowledge. *Nucleic acids research*, *47*(D1), D506-D515.

71. Van der Auwera, G. A., Carneiro, M. O., Hartl, C., Poplin, R., Del Angel, G., Levy-Moonshine, A., ... & DePristo, M. A. (2013). From FastQ data to high-confidence variant calls: the genome analysis toolkit best practices pipeline. *Current protocols in bioinformatics*, *43*(1), 11-10.

72. Wang, C. K., Han, P. L., Zhao, Y. W., Yu, J. Q., You, C. X., Hu, D. G., & Hao, Y. J. (2021). Genome-wide analysis of auxin response factor (ARF) genes and functional identification of MdARF2 reveals the involvement in the regulation of anthocyanin accumulation in apple. *New Zealand Journal of Crop and Horticultural Science*, *49*(2-3), 78-91.

73. Wang, J., & Zhang, Z. (2021). GAPIT Version 3: boosting power and accuracy for genomic association and prediction. *Genomics, proteomics & bioinformatics*, *19*(4), 629-640.

74. Wang, S., Chang, Y., & Ellis, B. (2016). Overview of OVATE FAMILY PROTEINS, a novel class of plant-specific growth regulators. *Frontiers in plant science*, *7*, 417.

75. Wang, S., Chang, Y., Guo, J., & Chen, J. G. (2007). Arabidopsis Ovate Family Protein 1 is a transcriptional repressor that suppresses cell elongation. *The Plant Journal*, *50*(5), 858-872.

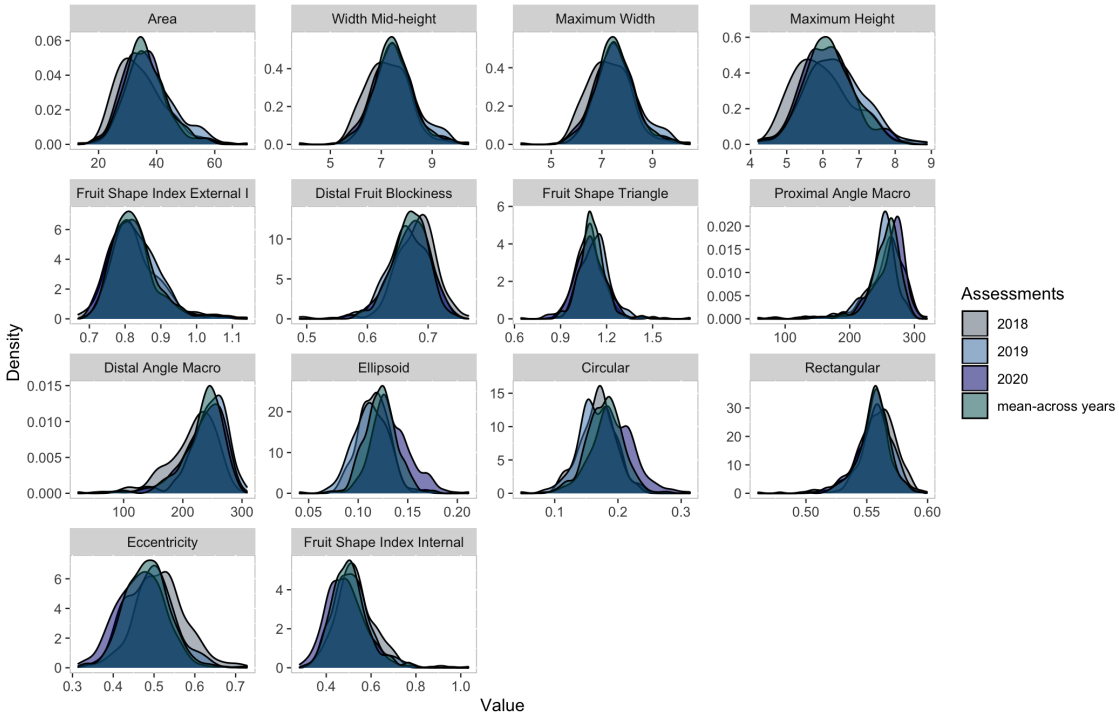
76. Wickham, H. (2016). Package 'ggplot2': elegant graphics for data analysis. *Springer-Verlag New York. doi*, *10*, 978-0.

77. Wolfe, D., Dudek, S., Ritchie, M. D., & Pendergrass, S. A. (2013). Visualizing genomic information across chromosomes with PhenoGram. *BioData mining*, *6*(1), 1-12.

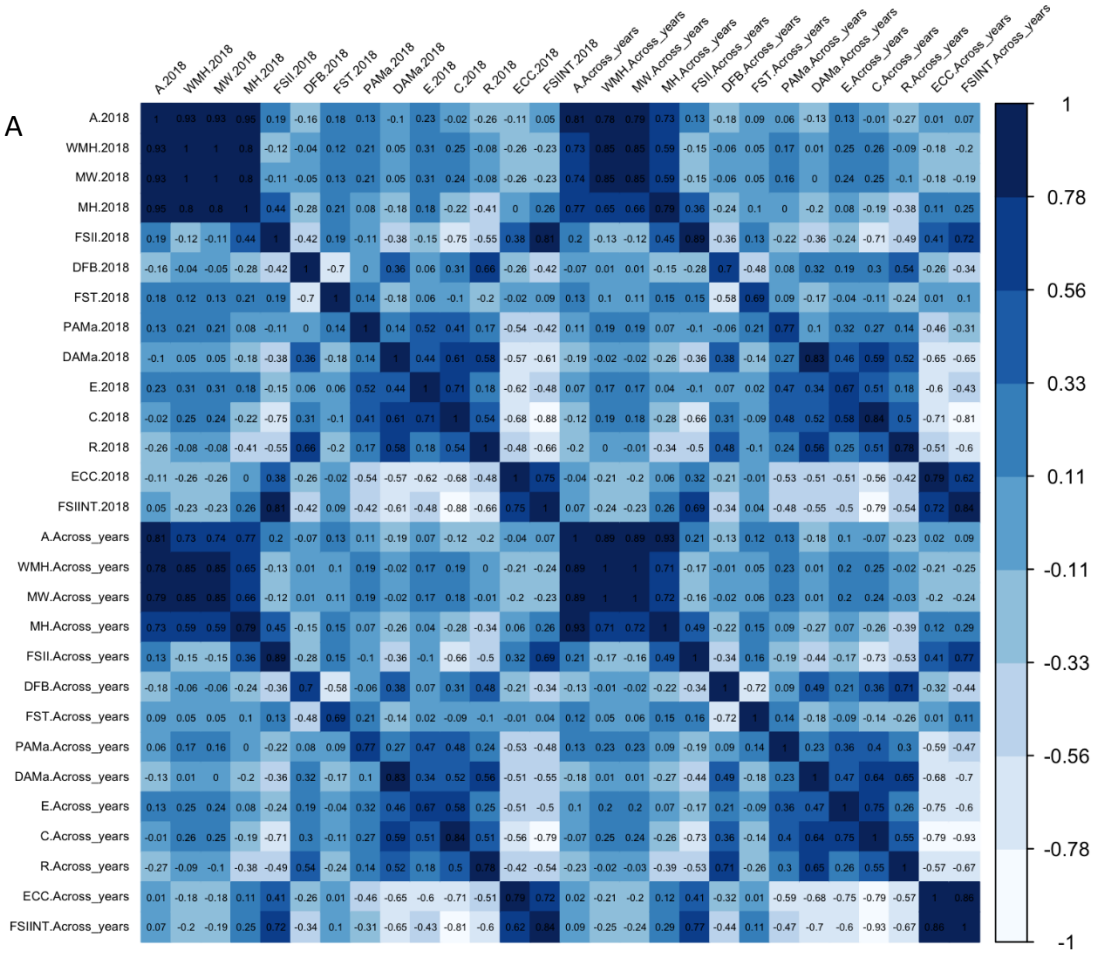
78. Wu, S., Clevenger, J. P., Sun, L., Visa, S., Kamiya, Y., Jikumaru, Y., ... & van der Knaap, E. (2015). The control of tomato fruit elongation orchestrated by sun, ovate and fs8. 1 in a wild relative of tomato. *Plant Science*, *238*, 95-104.

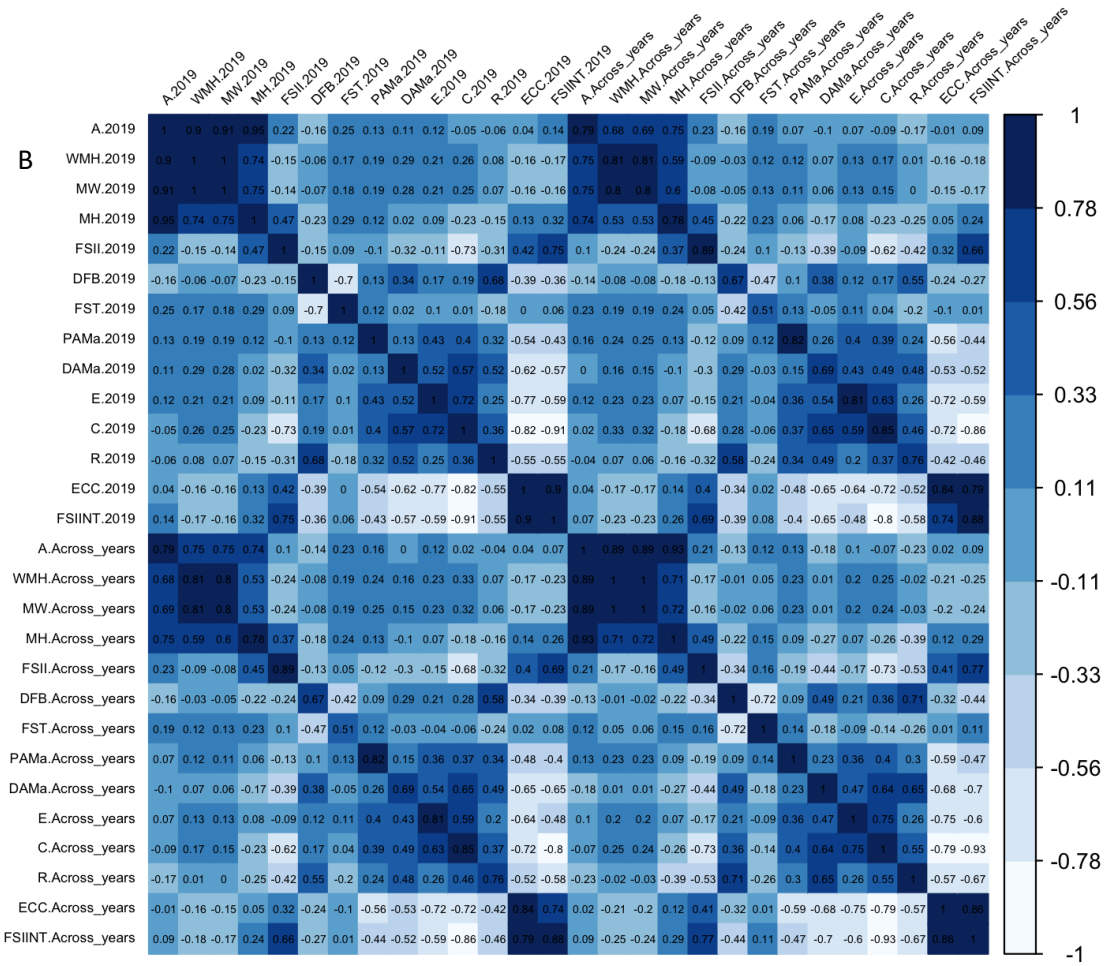
79. Yang, C., Ma, Y., He, Y., Tian, Z., & Li, J. (2018). Os OFP 19 modulates plant architecture by integrating the cell division pattern and brassinosteroid signaling. *The Plant Journal*, *93*(3), 489-501.
80. Yang, C., Shen, W., He, Y., Tian, Z., & Li, J. (2016). OVATE family protein 8 positively mediates brassinosteroid signaling through interacting with the GSK3-like kinase in rice. *PLoS Genetics*, *12*(6), e1006118.
81. Yao, J. L., Xu, J., Cornille, A., Tomes, S., Karunairetnam, S., Luo, Z., ... & Gleave, A. P. (2015). A micro RNA allele that emerged prior to apple domestication may underlie fruit size evolution. *The Plant Journal*, *84*(2), 417-427.
82. Yin, L. (2018). A high-quality drawing tool designed for Manhattan plot of genomic analysis.
83. Zhang, L., Hu, J., Han, X., Li, J., Gao, Y., Richards, C. M., ... & Cong, P. (2019). A high-quality apple genome assembly reveals the association of a retrotransposon and red fruit colour. *Nature communications*, *10*(1), 1-13.
84. Zhou, H., Ma, R., Gao, L., Zhang, J., Zhang, A., Zhang, X., ... & Han, Y. (2021). A 1.7-Mb chromosomal inversion downstream of a PpOFP1 gene is responsible for flat fruit shape in peach. *Plant biotechnology journal*, *19*(1), 192-205.
85. Zygier, S., Chaim, A. B., Efrati, A., Kaluzky, G., Borovsky, Y., & Paran, I. (2005). QTLs mapping for fruit size and shape in chromosomes 2 and 4 in pepper and a comparison of the pepper QTL map with that of tomato. *Theoretical and Applied Genetics*, *111*(3), 437-445.

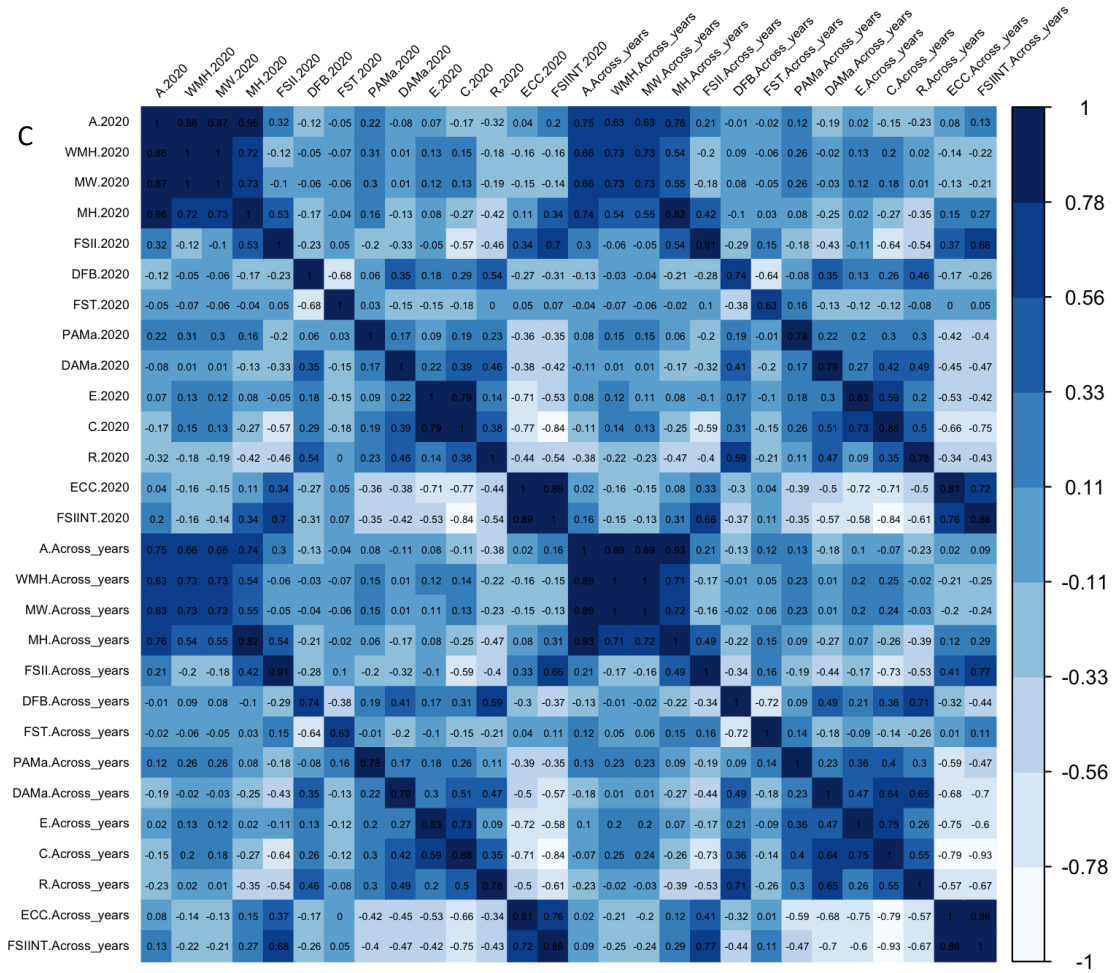
SUPPLEMENTARY MATERIAL



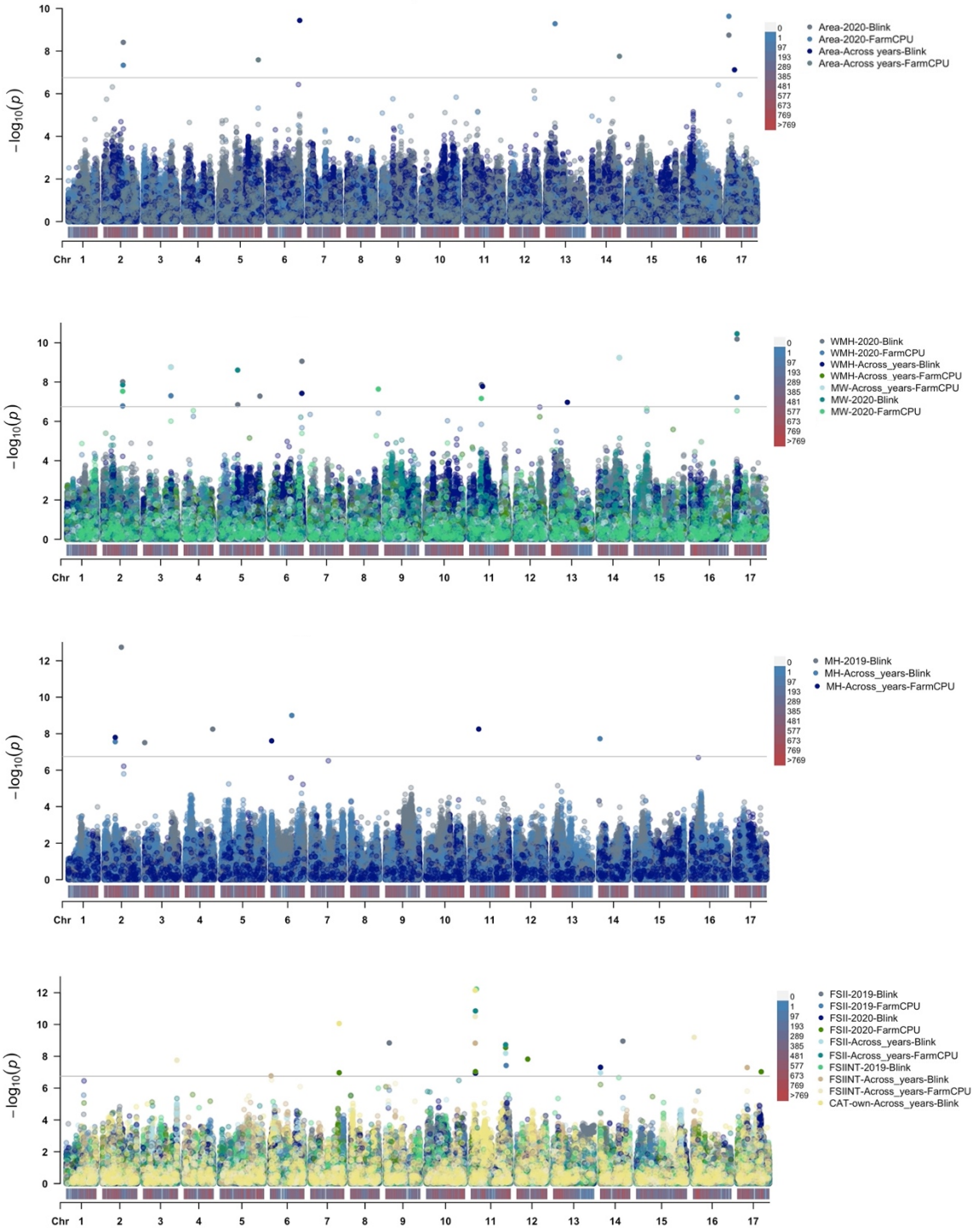
Supplementary Figure 2.1: Density plots of the distribution of the data corresponding to the years evaluated and mean-across years.

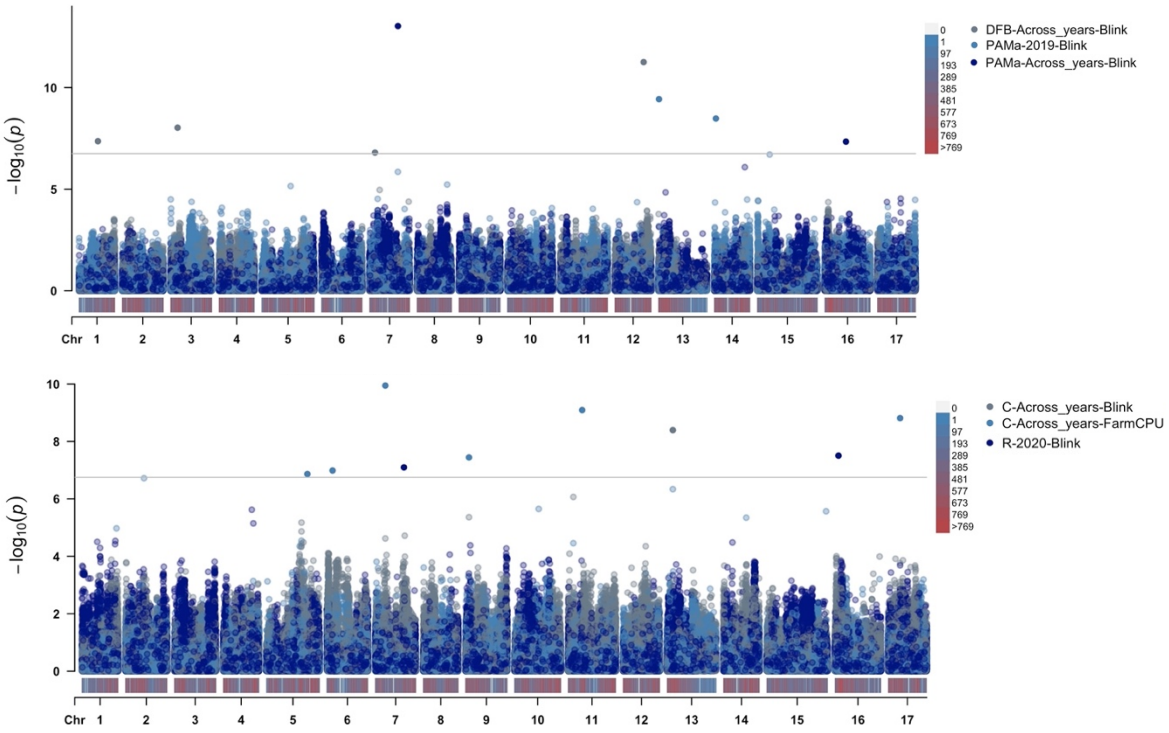



B
A.2019
WMH.2019
MW.2019
MH.2019
FSII.2019
DFB.2019
FST.2019
PAMa.2019
DAMa.2019
E.2019
C.2019
R.2019
ECC.2019
FSIINT.2019
A.Across_years
WMH.Across_years
MW.Across_years
MH.Across_years
FSII.Across_years
DFB.Across_years
FST.Across_years
PAMa.Across_years
DAMa.Across_years
E.Across_years
C.Across_years
R.Across_years
ECC.Across_years
FSIINT.Across_years
1
0.78
0.56
0.33
0.11
-0.11
-0.33
-0.56
-0.78
-1

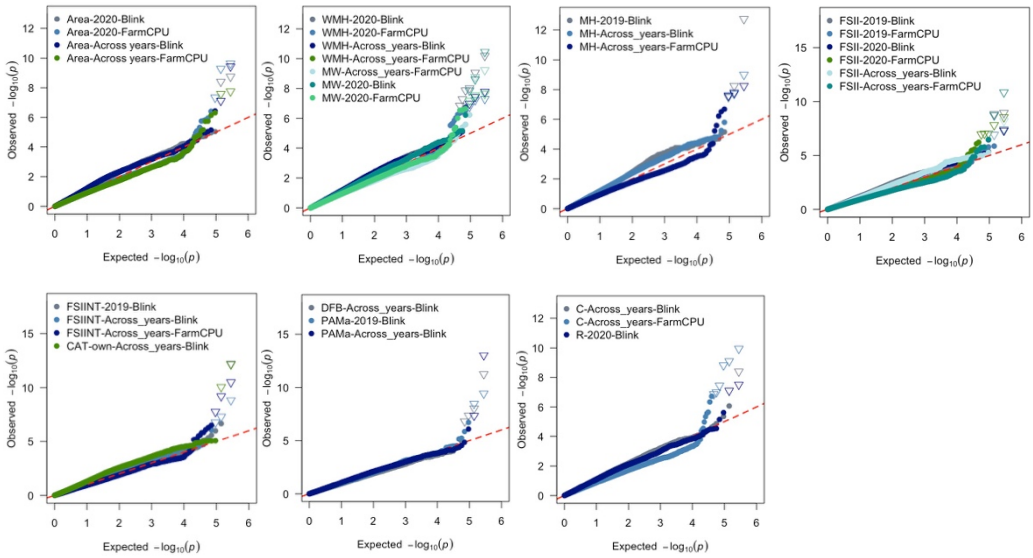

C
A.2020
WMH.2020
MW.2020
MH.2020
FSII.2020
DFB.2020
FST.2020
PAMa.2020
DAMa.2020
E.2020
C.2020
R.2020
ECC.2020
FSIINT.2020
A.Across_years
WMH.Across_years
MW.Across_years
MH.Across_years
FSII.Across_years
DFB.Across_years
FST.Across_years
PAMa.Across_years
DAMa.Across_years
E.Across_years
C.Across_years
R.Across_years
ECC.Across_years
FSIINT.Across_years
1
0.78
0.56
0.33
0.11
-0.11
-0.33
-0.56
-0.78
-1

Supplementary Figure 2.2: Spearman correlation between year assessment and mean across-years of apple fruit measurements. A, Corresponding to 2018 with mean across-years. B, Corresponding to 2019 with mean across-years. C, Corresponding to 2020 with mean across-years. Inside each square indicating the correlation score. The acronymous corresponding to Area (A), Width Mid-height (WMH), Maximum Width (MW), Maximum Height (MH) Fruit shape index external I (FSII), Distal fruit blockiness (DFB), Fruit shape triangle (FST), Proximal angle macro (PAMa), Distal angle macro (DAMa), Ellipsoid (E), Circular (C), Rectangular (R), Eccentricity (ECC) and Fruit shape internal (FSIINT).

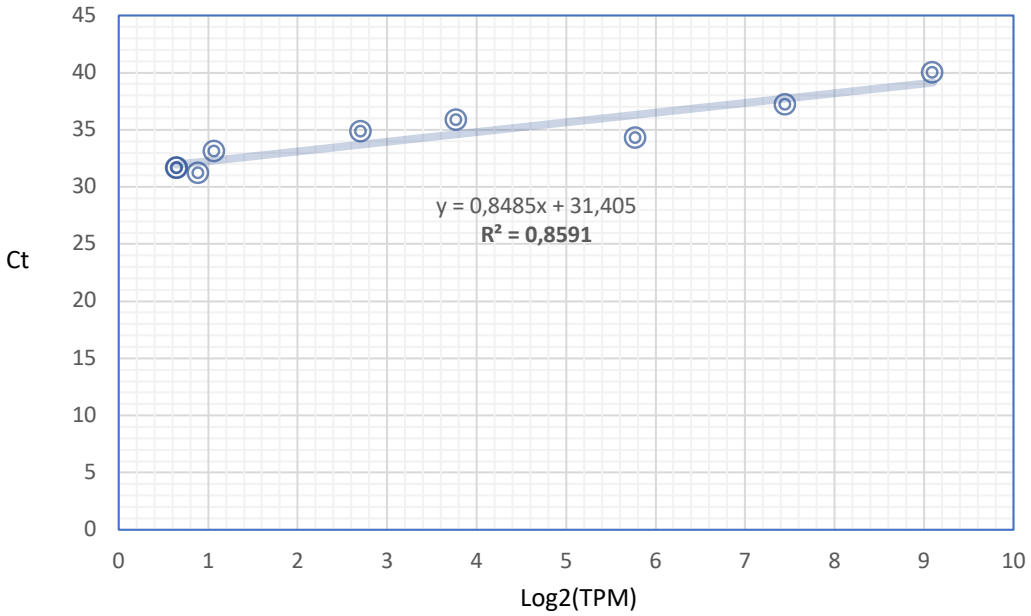




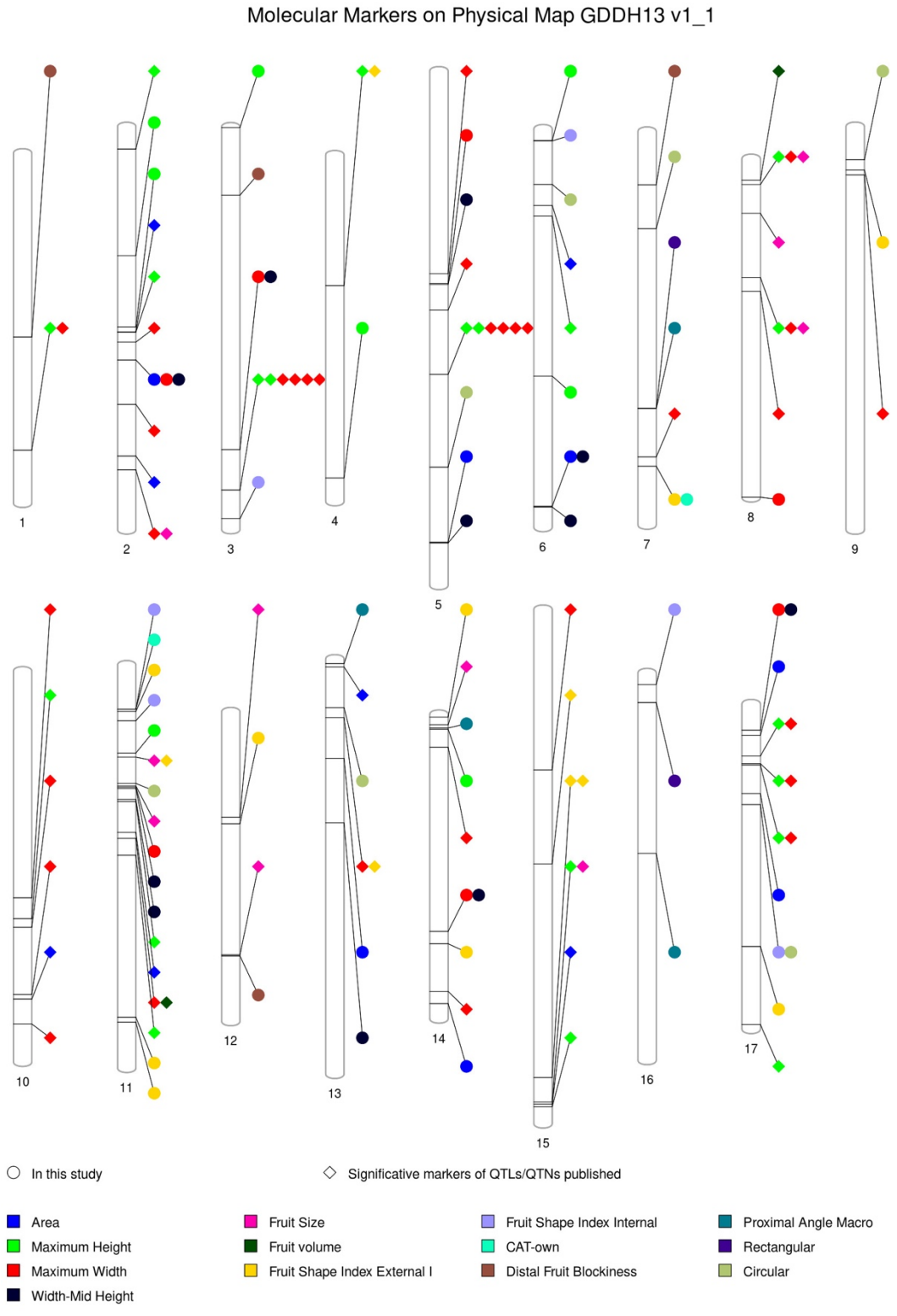
Supplementary Figure 2.3: Manhattan plot of GWAS results on phenotype per year or across years and GWAS model. The threshold (6,751) is represented by the gray line. The bar in the X axe represents the density of SNPs per chromosome, from gray to red (lower to higher). The legends of each plot correspond to the trait-year or across years-model.



Supplementary Figure 2.4: QQplot of GWAS results on mean phenotype per year or across years and GWAS model. The legends of each plot correspond to the trait-year or across years-model.



Supplementary Figure 2.5: Validation of RNA-seq data by qPCR of the HF43536 gene. The trend line is represented by the equation below and R-squared. Log2(TPM) on the x-axis and Ct on the y-axis.



Supplementary Figure 2.6: Molecular Markers on Physical map GDDH13v1.1. Significant markers for mapping in apple fruit measures, including markers published described in some QTLs/QTNs analysis and QTNs for this study. Symbols: circle, correspond to "in this study" and diagonal, "significative markers of QTLs/QTNs published". Each color corresponds to different trait for size and shape. See more details in **Supplementary Table 2.7**.

CHAPTER 3

Genetic study of fruit shape along apple development from a morphologic, histologic, and differential gene expression perspective, in three fruit shape typologies.

Dujak, Christian[1], Garcia, Beatriz[1] and Aranzana, Maria José[1,2]

[1] Centre for Research in Agricultural Genomics (CRAG) CSIC-IRTA-UAB-UB, Campus UAB, 08193, Bellaterra, Barcelona, Spain.

[2] IRTA (Institut de Recerca i Tecnologia Agroalimentàries), 08140, Caldes de Montbui, Barcelona, Spain.

ABSTRACT

Pome fruits consist of two main fleshy tissues, the core and the cortex. The core derives from the ovary, while the cortex (also known as the hypanthium) derives from the fusion of accessory tissues, including sepals, petals and stamens. The cortex contributes to more than 70% of the volume of the mature fruit. Fruit growth and development starts at fruit set. During the first weeks, fruit grow due to intensive cell division which is followed by linear growth mainly due to cell expansion, that can occur till ripening. While fruit growth along development has been deeply studied, the evolution of fruit shape and the genomic mechanisms regulating the process has been poorly understood. Here we studied histological cuts of the parenchyma of the hypanthium of fruits of three cultivars (one flat, one round and one oblong) along development. Cell parameters (number of cells, cell area and intercellular space area) were evaluated in fruits of the three genotypes at 0, 61 and 98 days after anthesis (DAA). Total RNA sequencing served to identify differentially expressed genes (DEG) along development (time-course differentially expression analysis, TC-DEA) and between fruits of different shape. The TC-DEA identified multiple genes differentially expressed along development within and between cultivars. Some of the ones explaining more than 50% of the variance were phytohormones previously described to have a role in fruit development. Some of the genes identified in the GWAS analysis (Chapter 2) were found differentially expressed in the contrasts studied. Among them the *MdOFP4*, which was not expressed in fruits of the oblong cultivar. Whole genome DNA sequence revealed a deletion in the promoter of the gene. Further analyses are required to validate the association of this polymorphism for its use in breeding.

Keywords: Apple fruit growth, fruit development, histological analysis, time-course analysis, differentially expressed genes.

INTRODUCTION

Apple (*Malus* × *domestica* L. Borkh) fruit development has been the subject of several studies. After ovule fertilization, the seeds begin to develop in the ovary and increase mitotic division from the central part of the sac to the margins of the ovary, this phase is known as fruit set. The second phase is fruit development, which consists of four stages: cell division, cell expansion, ripening and maturation (Janssen et al., 2008). For each phase, multiple molecular pathways have been described that trigger gene expression responsible for development, as well as genes that have not yet been identified.

During fruit development cell production increases exponentially, *cyclin-dependent kinases* (*CDKs*) and *cyclins* (*CYC*) genes are involved in cell proliferation and their regulation during the phases of mitotic division. During the fruit growth process there are stages of cell division and subsequent cell elongation (Inze and De Veylder, 2006; Malladi and Johnson, 2011). Phytohormones also play an important role in fruit development. In tomato, auxin (AUX) in the early fruit stage promotes cell division and gibberellin (GA) promotes cell expansion (Serrati et al., 2007). In apple, AUX has also been shown to naturally increase cell division in the cortex (Devoghalaere et al., 2012), and GA increases cell number and size and have an effect on fruit symmetry, as found after localized applications (Nakagama et al., 1967). In dicotyledonous plants, the "ABC model" describes the interaction of homeotic genes to ultimately establish the identity

of the basic organs. Post-fertilization, some of the fruit tissues derive from floral organs (Ma and dePamphilis, 2000).

In apple, several genes are expressed in the floral organ tissues, such as the *MADs-Box* genes (transcription factors), the *APETALA2* (*AP2*) gene in the sepals (Yao et al., 1999) , as well as in the cortex/flower tube during early fruit development (Kotoda et al., 2000). In an apple, a microRNA (miR172) that regulates the *AP2* gene has an effect fruit size (Yao et al., 2015). Another gene identified in petal tissue was the *PISTILLATA* (*PI*) (Yao et al., 2001), in transgenic plant trials the gene was found to influence fruit shape with a flat appearance (Yao et al., 2018).

In tomato, a model species for the study of fleshy fruits, cellular and molecular mechanisms that control fruit size and shape have been identified by numerous research studies (Mauxion et al., 2021). Among them, *SUN, OVATE and FS8.1* have been described as genes/QTL responsible of controlling the ovary and fruit elongation (Wu et al., 2015; Wang et al., 2019; Mauxion et al., 2021).

The objective of this work is to gain knowledge on apple development in fruits with the three most representative shape classes: oblate, spheroid and oblong. The analyses include monitoring at 10 points during growth acquiring morphometric parameters, the histologic study of the parenchyma of the hypanthium at thre developmental stages, coupled with the study of RNA and DNA high-throughput sequences. These results provide further insight into the physiology of fruit development and on the genes involved in apple fruit shape determination.

MATERIALS AND METHODS

Plant material

Nine genotypes were selected from the REFPOP apple collection (Jung et al., 2020) located in Lleida (Spain) based on the visual classification (CAT-own) used in **Chapter 1**, as Oblate or flat ('Carrata', 'Grand'mere' and 'Gros Api'), Spheroid or round ('Kansas Queen', 'Horei' and 'Pero dourado') and Oblong ('Skovfoged', 'Giambun' and '12_O063'). At least 3 samples were collected per cultivar at 0, 13, 23, 35, 47, 61, 70, 84, 98 days after anthesis (DAA) as well as at harvest.

Height, width and FSI measurements were obtained at each collection point.

A Kruskal-Wallis test was conducted to compare the measures per cultivar at each collection point and plotted on dotted lines using ggplot2 package (Wickham, 2016) in R Core Team (2022) program.

Tissue processing, Embedding and sectioning

For tissue processing, three points along fruit development (0, 61 and 98 DAA) and three genotypes ('Grand'mere', 'Kansas Queen' and 'Skovfoged') were selected, each one representing a fruit shape (Oblate, Spheroid and Oblong, respectively).

At least 2 replicates of each genotype and point were processed, cutting longitudinally from the central section of each flower and the fruit hypanthium, picking small portions of sample (from the skin to the core) from the widest part. These samples were fixed in FAA solution (Formaline-Acetic Acid-Ethanol 70%) for 72hs, followed by paraffin embedding that consisted in several washes with distilled water, immersion in 3N HCl for 30 min, successive changes with ethanol from 50% to 100%, xylene-paraffin embedding in three proportions (3:1, 1:1, 1:3) and finally three changes of 100% paraffin. Each change of solution was made at 60ºC for 2-3 hours. Once the blocks were

made, were cut in 10µm thick sections using LEICA® RM2125RT microtome. Later they were deparaffinized and dyed with Safranin-Fast Green for visualization of the cells in the OLYMPUS® optical microscope. Measurements were cell area, cell number and intercellular space area within a circumference of 0.5 mm^2 for all samples and points (0, 61 and 98 DAA). In the case of the 61 and 98 DAA samples, cell parameteres were measured at 5 positions along the transversal axis (from epidermis to core) for both side left and right of the longitudinal sections of the fruit. For the analysis of the cells, all observations were saved in images and measured manually using ImageJ-Fiji program (Schindelin et al., 2012). Several statistical tests were performed, such as Shapiro-Wilk (nomality test), Kruskal-Wallis and Dunn with the adjusted p-values for Holm or Bonferroni correction (comparison between groups), and were plotted in Boxplot-Violin detailing p-value values using *ggstatsplot* package (Patil, 2021) in the R Core Team program (2022).

DNA and RNA extraction

For DNA extraction, leaf samples were collected from three branches in two clones, and for total RNA extraction fruit were processed from points 13, 61 and 98 (DAA) of the varieties 'Grand'mere' (GRA), 'Kansas Quee'n (KAN), and 'Skovfoged' (SKO) with their three biological replicates.

Total DNA for whole genome sequencing (WGS) and whole genome bisulphite sequencing (WGBS) was extracted with a CTAB modified protocol (Inglis et al., 2018). Total RNA was extracted with Maxwell® RSC simplyRNA tissue kit, using Maxwell® RSC instrument. RNA quality and quantify was evaluated with Bioanalyzer® and sent to Novogene (London, England).

Sequencing libraries filtering, mapping to reference genome and quality control.

The whole genome DNA, DNA bisulfite-treated and mRNA sequencing libraries were filtered by sequencing quality (reads with a Phred score < 30 were removed), and remaining Illumina sequencing adaptors were trimmed by using Trim-Galore version 0.6.1 (Krueger et al., 2015). Burrows-Wheeler Aligner (BWA-MEM) version 0.7.17 (Li and Durbin, 2009) was used to map clean DNA reads to apple anther-derived homozygous genotype 'Hanfu' ('Dongguang' × 'Fuji') genome HFTH1 (Zhang et al., 2019). While bisulfite mapper and methylation caller Bismark software version 0.22.3 (Krueger & Andrews, 2011) was used to perform alignments of bisulphite-treated DNA reads. In the particular case of WGBS, the reference genome was converted into a bisulphite version (C to T and G to A converted). Besides, sequence reads were also transformed into fully bisulphite-converted versions before reads were aligned to similarly converted versions of the genome in a non-directional manner. High-quality RNA sequencing libraries were also mapped to HFTH1 by using HISAT version 2.1.0 (Kim et al., 2015) with default settings parameters. The reference genome (converted and unconverted for bisulphyte-treated DNA mapping) were previuosly indexed with Bowtie version 2.3.4.1 (Langmead et al., 2012).

Statistical data of mapped and unmapped reads in Binary Alignment Map (BAM) files were analyzed using SAMStat version 1.5.1 (Lassman et al., 2011). The index used to determine the quality of alignment and assembly was the Mapping Quality Score (MAPQ) (Li et al., 2008), which quantifies the probability of a misplaced read. Assembled libraries in BAM format were filtered by MAPQ score, based on the quality of the alignment. Multiple aligned reads (reads with MAPQ< 30 or not properly aligned according to the mapper) were filtered, and statistic reports were obtained with

Samtools version 1.9 (Li et al., 2009). Besides, Samtools was used to transform, index, and sort the files generated by the mappers according to the protocol's needs.

Quality control reports were obtained before and after filtering and mapping with FastQC version 0.11.5 (Andrew, 2010) to ensure high-quality standards for downstream analyses. All the quality reports were summarized in an Html file by using MultiQC version 1.9 (Ewels et al., 2016).

Differentially expressed genes and differentially methylated region analyses.

To identify differentially expressed genes, transcript quantification and count matrix construction were performed with *featureCounts* (Liao et al., 2014) setting the parameters to paired-end sequencing, avoiding chimeric count fragments (those fragments that have their two ends aligned to different chromosomes), specifying as feature exon feature type for reading counting and annotated as transcript and allowing overlapping features for the differential use of exon during alternative splicing. The results obtained were normalized to Transcript Per Million (TPM) (Li and Durbin, 2010). The batch effect was checked by sva R package version 3.12 (Leek et al., 2012). In addition, preliminary exploratory analysis and visualization of the samples from the dataset were performed. For count matrix normalization, a regularized-logarithm transformation (*rlog*) was applied, recommended to stabilize the variance across the mean for negative binomial data with a dispersion-mean trend and a low number of samples ($n < 30$) (Love et al., 2014).

The time course differential expression analysis (TC-DEA) was done with by DESeq2 version 1.30.0 (Love et al., 2014), which can be used to analyze time course experiments, finding those genes that react in a condition-specific manner over time by using Likehood Ratio Test (LRT) and removing the interaction factor time:variety from the

model. Previously to DEGs analysis, the count matrix were pre-filtered to remove all the genes with less than 10 counts in almost 3 of the samples to avoid genes that have very low or no expression. Once DEGs analysis was run, genes with false discovery rate (FDR) under 0.05 were filtered (Benjamini and Hochberg, 1995). The consensus result from time course analysis was represented as Venn diagrams done with the web application InteractiVenn (Heberle et al., 2015).

Correlation and R-squared

Spearman's correlation and coefficient of determination (R-squared) between TPM of the annotated TC-DEA genes and phenotypic data (FSI, Height and Width) and cell data (cell area, cell number and intercellular spaces) were calculated. In the case of correlation, all genes annotated in the TC-DEA and visualized in histograms were selected for further analysis. candidate genes from the GWAS results (**Chapter 2**) and annotated genes from the TC-DEA with R^2 >0.5 were filtered out, and were plotted in dot and line plots. In addition, the comparison of gene expression in TPM between varieties and DAA was carried out by Kruskal-Wallis test with $p<0.05$ and plotted in line plot. All plots were made using ggplot2 package (Wickham, 2016) in R Core Team (2022) program.

RNA-seq validation was performed in **Chapter 2** (see details in **Supplementaries Table 2.6 and Figure 2.5**)

Differential expression of genes between genotypes and at the three points DAA was evaluated by testing the distribution of the data with Shapiro test, Anova-one way, and Kruskal-Wallis for normal and non-normal distributed data, respectively. Confidence level was set at $p<0.05$. Differences between genotypes were determined with Tukey HSD test, using the normalized count matrix in TPM.

Differentially methylated regions (DMR) were estimated with the R package DSS version 2.40 (Wu et al., 2013). The parameters for DML assignment were set for any absolute differences in methylation levels between groups (delta = 0) and statistical threshold significance of *p*-value < 0.00005. The spanning smoothing was activated in a range of 200 cytosines methylated. For DMR identification the parameters were set to a minimum of three methylated cytosines in a length of 50 pb and a minimum of 5% of methylated sites in the region was required for a statistically significance. Finally, near DMR were merged when the distance between then where under 100 pb.

Gene annotation

DEGs were annotated based on previously annotation of HFTH1 genome *(Zhang et al., 2019)* in Gene Onthology (GO) terms (Ashburner et al., 2000), InterPro (IPR) (Hunter et al., 2009), Kyoto Encyclopedia of genes and genomes (KEGG orthologs and pathways) (Kanehisa et al., 2016), non-redundant proteins sequences from NCBI (RefSeq) (Pruitt et al., 2007), *Arabidopsis thaliana* orthologs from the Arabidopsis Information Resource (TAIR) (Lamesch et al., 2012) and computer-annotated protein sequence database, from the translation of coding sequences (UniProtKB/TrEMBL) (Uniprot Consortium, 2019).

Variants calling

Local alignment and variants calling (SNPs and Indels) were conducted by using Genome Analysis Toolkit (GATK) version 4.1.6 (Van der Auwera et al., 2013). SNPs with a minimum Phred-scaled confidence threshold of 30 were removed. The potential false-positive variants were avoided by performing a hard filtering over SNPs setting the parameters of variant confidence (QUAL > 30) normalized by unfiltered allele depth of variants samples (QD < 2.0), allele-specific strand bias estimated using Fisher's Exact Test (FS > 60.0), strand odds ratio (SOR > 3), root mean square mapping quality over all reads at

the site (MQ < 40.0), u-based z-approximation from the Rank Sum test for mapping qualities (MQRankSum < –12.5), u-based z-approximation from the Rank Sum Test for site position within reads (ReadPosRankSum < –8.0), Phred-scaled p-value for exact test of excess heterozygosity (ExcessHet > 54.69), total depth (DP < 30) and the Phred-scaled confidence that the genotype assignment (GT) is correct (GQ > 15).

The IGV program (Thorvaldsdóttir et al., 2011) was used to visualize the methylated regions in CHG, CHH and CpG, adding the SNPs variants and reads sequences of the varieties (GRA, KAN and SKO).

RESULTS

Shape and size attributes along fruit development

Fruits of nine genotypes of selected according to their fruit shape observed in previous harvest years, such as oblate or flat ('Carrata', 'Grand'mere' and 'Gros api'), Spheroid or round ('Kansas Queen', 'Horei' and 'Pero Dourado'), and oblong ('Skovfoged', 'Giambun' and '12_O063') were sampled from March to October 2019, 0 days after DAA anthesis until fruit harvest (**Supplementary Table 3.1**).

The FSI values along fruit development separated the three groups (oblate, spheroid and oblong) in the nine genotypes, while other size attributes as height, width and weight did not (**Figure 3.1**). The FSI provided the criteria for selecting one representative genotype per shape (**Figure 3.1** and **Supplementary Data 3.1**). Fruits from 'Grand'mere' were the largest, with outstanding width and weight. In contrast, the "Gros api" genotype was the smaller, with lower height and weight values. (**Figure 3.1** and **Supplementary Data 3.1**).

Fruits of three of the nine genotypes, 'Grand`mere' (GRA), 'Kansas Queen' (KAN) and 'Skovfoged'(SKO), at 0, 61 and 98 DAA, were selected for further study (**Figure 3.2**). FSI

values of the fruits of each genotype at the three developmental stages were significantly different (**Supplementary Figure 3.1**).

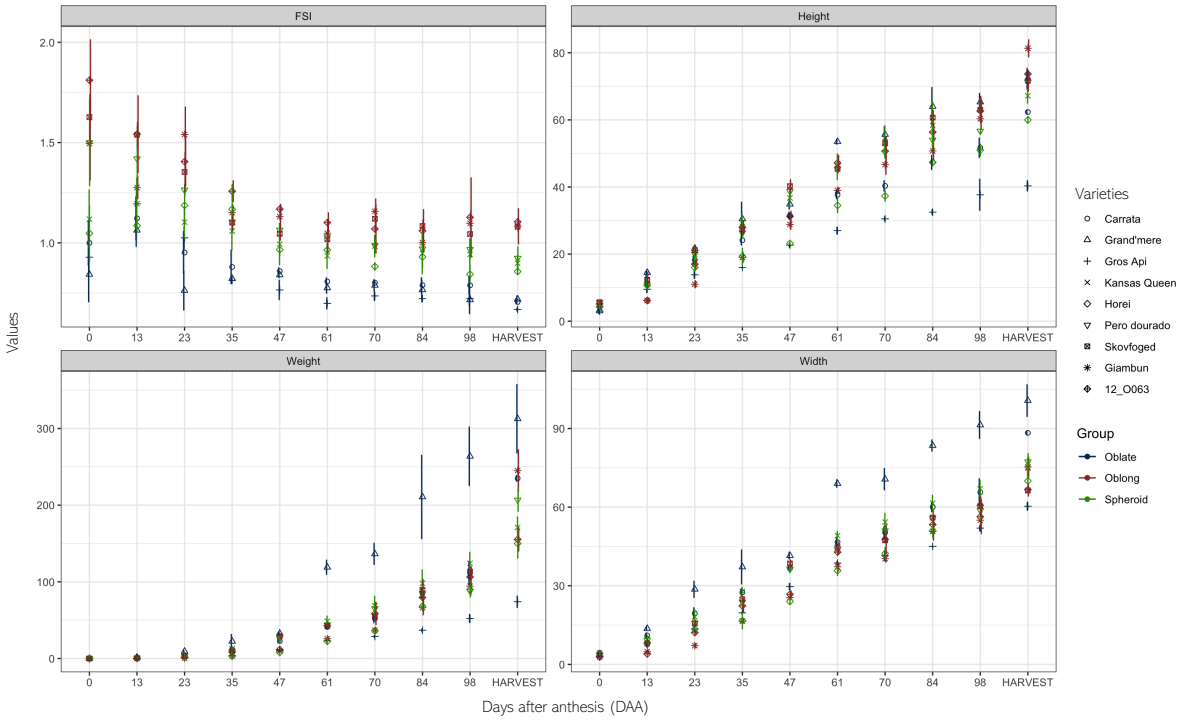


Figure 3.1: Dot plot of the growth from days after anthesis to fruit harvest. Samples were collected at 10 points along the development, nine varieties representing three shape types were selected, Oblate ('Carrata', 'Grand'mere' and 'Gros Api'), Spheroid ('Kansas Queen', 'Horei' and 'Pero dourado'), Oblong ('Skovfoged', 'Giambun' and 12_O063). The measurements were FSI (height/width), height(mm), weight (grams) and width (mm).

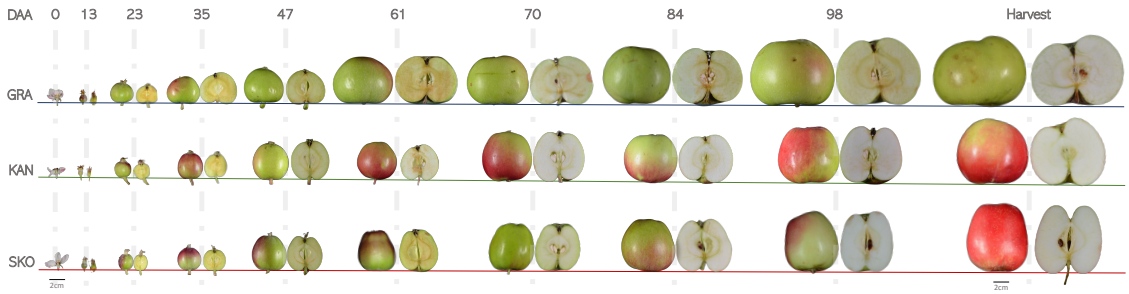


Figure 3.2: Images of fruit development in three apple shape types, corresponding to three varieties GRA ('Grand'mere'), KAN ('Kansas Queen') and SKO ('Skovfoged'), collected along development from stage 0 DAA (Days after anthesis) to harvest. Scale bar 2cm.

Parenchyma cells and intercellular spaces along fruit development

Cuts of parenchyma of GRA, KAN and SKO fruits at the three points of development (0, 61 and 98 DAAA) were observed at the microscope as shown in **Figure 3.3** and **Supplementary Data 3.2**.

At the 0 DAA point, longitudinal cuts showed differences in the dimensions and shape of the ovary between genotypes, for example, in GRA and SKO the maximum diameter is similar, but the apical part of the ovary in SKO is more elongated while in GRA is shorter and wider. Regarding the conformation of the parenchyma, significant differences between genotypes were found in the cell area, cell number and intercellular spaces (IS). At the 0 DAA point, them mean cell area value for KAN was 2.50E-04mm^2, 2.88E-04mm^2 for GRA and 3.99E-04mm^2 for SKO.

The cell number per field (0.5 mm^2), was also significantly different between KAN and SKO genotypes. The average cell number per genotype was lower in SKO with 1252 cells, followed by GRA with 1880 cells and by KAN with an average of 2105 cells. The intercellular spaces (IS) were also measured in the same analyzed fields. GRA tissue did not have IS at the 0 DAA point, while KAN and SKO had the same IS average value of 0.04mm^2. An important aspect to highlight at the 0 DAA point, according to the correlation values between these variables, was that when cell area increases, cell count decreases (**Figure 3.3, 3.4** and **Supplementary Data 3.2**).

Along fruit development, and in particular at 61 and 98 DAA, fruit growth differentiates to ultimately acquire the final shape and size. At these stages, the cells evaluated corresponded to 5 positions along the axis of hypanthium area, between the epidermis and the nucleus.

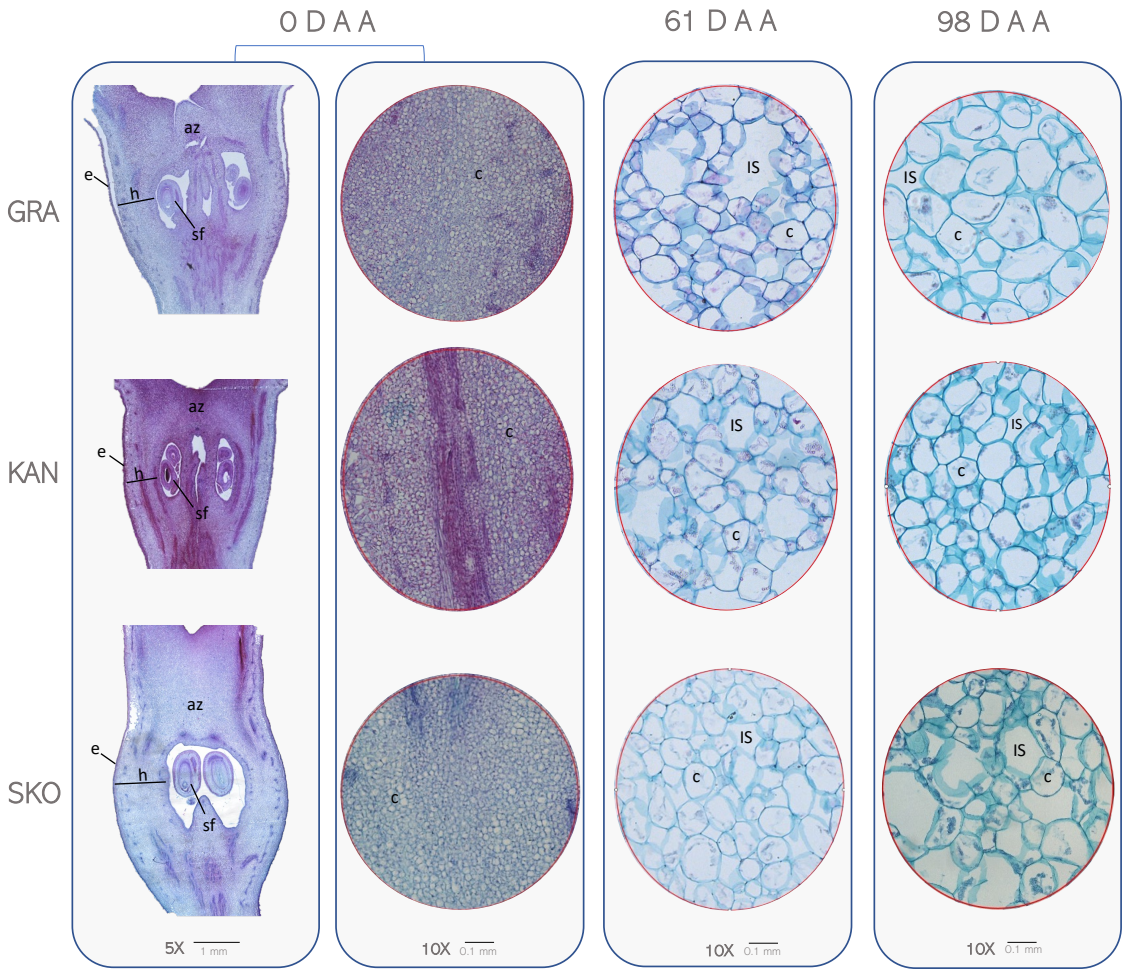


Figure 3.3: Microscopic images of longitudinal sections of flower and fruit at stages O, 61 and 98 DAA of three varieties GRA ('Grand'mere'), KAN ('Kansas Queen') and SKO ('Skovfoged'). In each circle (0.5 mm^2) cells of the parenchyma tissue deriving from the ovary and posteriori hypanthium area are observed. Letters, **e** (epidermal layer), **h** (cells of the ovary, posteriorly hypanthium area), **sf** (seed formation), **az** (apical zone of the ovary), **c** (cell) and **IS** (intercellular spaces).

Taking the five positions per genotype and the three genotypes, at point 61 DAA the average cell area was 7.18E-03 mm^2. Differences in cell area were statistically significant between GRA and KAN, having GRA the smaller cells. The average cell count, when considering all genotypes, was 54.16 cells per field, with differences between KAN and SKO. SKO had the higher number of cells. Observing the SI data have increased exponentially from 0 to 61DAA in the three cultivars, considering the differences between the GRA and SKO genotypes (**Figure 3.3, 3.4** and **Supplementary Data 3.2**).

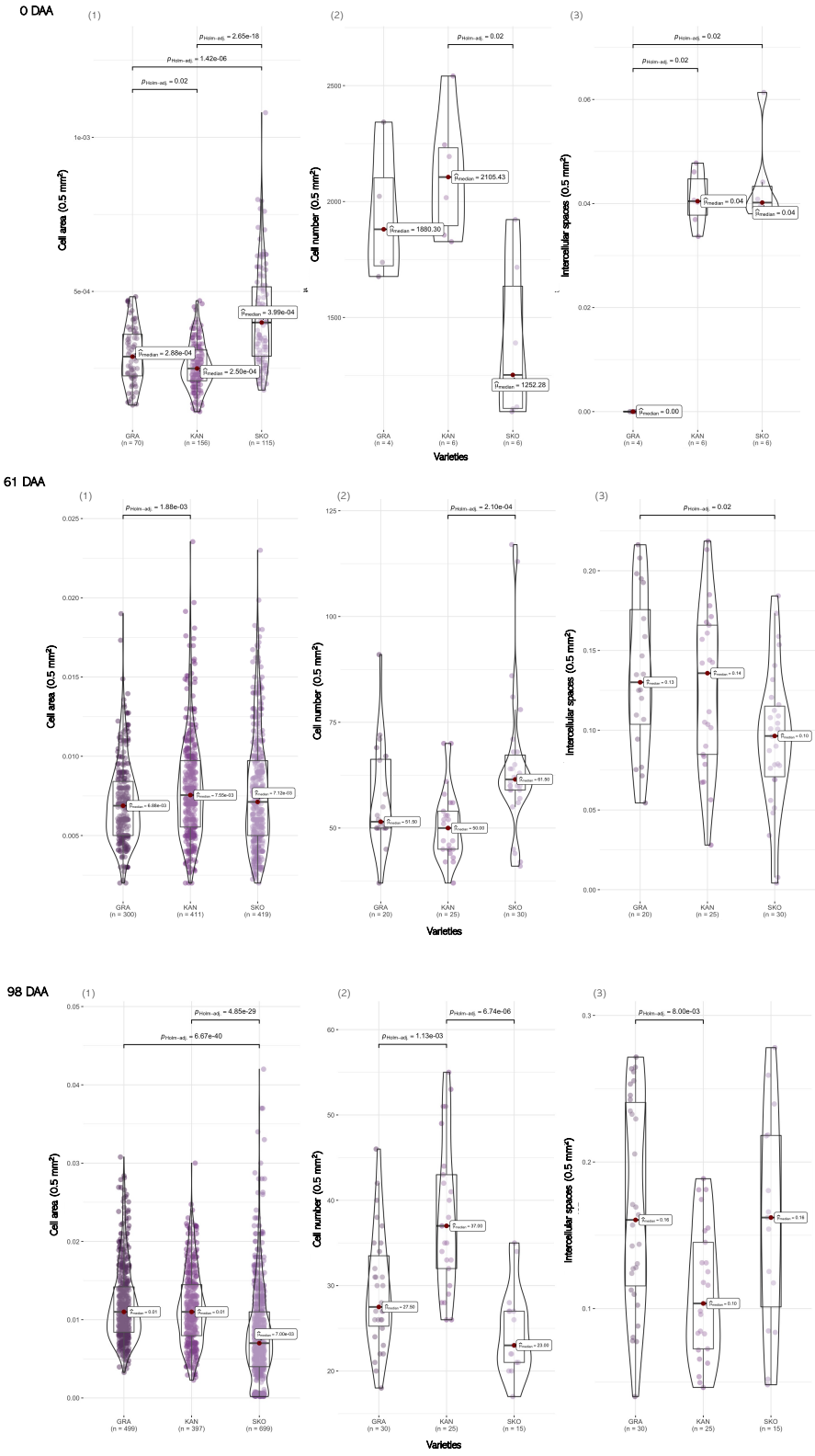


Figure 3.4: Boxplot of parenchyma tissue analysis in the hypanthium area by longitudinal sections of three apple varieties, 'Grand'mere' (GRA), 'Kansas Queen' (KAN) and 'Skovfoged' (SKO) taking measures such as cell area (0.5 mm^2), cell number (0.5 mm^2) and intercellular spaces (0.5 mm^2). Statistical data: **ODAA,** (1) $\chi 2\text{Kruskal} - \text{Wallis} = 79.18, \rho = 6.40\text{e} - 18, \text{n obs} = 341,$ (2) $\chi 2\text{Kruskal} - \text{Wallis} = 7.81, \rho = 0.02, \text{n obs} = 16$ (3)$\chi 2\text{Kruskal} - \text{Wallis} = 8.66, \rho = 0.01, \text{n obs} = 16.$ **61DAA,** (1) $\chi 2\text{Kruskal} - \text{Wallis} = 11.71, \rho = 2.86\text{e} - 03, \text{n obs} = 1130,$ (2) $\chi 2\text{Kruskal} -$

$\text{Wallis} = 15.89, \rho = 3.55\text{e}-04, \text{n obs} = 75$, (3) $\chi 2\text{Kruskal} - \text{Wallis} = 8.60, \rho = 0.01, \text{n obs} = 75$. **98DAA,** (1) $\chi 2\text{Kruskal} - \text{Wallis} = 220.49, \rho = 1.32\text{e}-04, \text{n obs} = 1595$, (2) $\chi 2\text{Kruskal} - \text{Wallis} = 24.53, \rho = 4.72\text{e}-06, \text{n obs} = 70$, (3) $\chi 2\text{Kruskal} - \text{Wallis} = 9.63, \rho = 8.11\text{e}-3, \text{n obs} = 70$.

We also compared the parameters at the five positions evaluated. Labeling them from 1 to 5 going to the external to the internal part of the fruit (i.e. position 1 is close to the epidermis and position 5 to the nucleus), only position 4 showed differences between the three genotypes for cell area. Regarding the cell count, positions 2, 3 and 4 showed differences for certain pairs of genotypes. However, we did no find differences for IS values between cultivars (**Supplementary Figure 3.2A-C**).

At 98 DAA, the fruit already has a shape close to the final one and maturation takes place. The average cell area increases to ~0.01 mm2, with differences observed in the SKO genotype due to its smaller cell area. At this stage, the number of cells per field decreased, with an average of ~29 cells. KAN was the genotype with the highest cell count. The average IS also increased, identifying differences between GRA and KAN (**Figure 3.3** and **3.4**).

When considering the positions of the fields in the fruit, in position 1 GRA had the larger cells while in position 2 it was SKO. Regarding the cell count, in positions 1, 4 and 5 the genotype KAN had higher numbers than the other two cultivars. In the case of IS measure, we found no statistically significant differences (**Supplementary Figure 3.2D-F**).

A correlation matrix between growth measurements (FSI, height, width) and cell analysis (cell area, cell number and IS) explained a positive correlation between cell area and IS with height, weight, and width, as well as the negative correlations of the cell number with height, weight, width and cell area (**Supplementary Figure 3.3**).

Differential gene expression analysis

Exhaustive filtration and trimming of sequencing and mapping libraries avoided the assignment of reads to chimaera genes, fragmented genes, duplicated regions, multiple mapping, aligned as singletons, aligned to secondary sequences, divided, or ambiguously assigned (**Supplementary Table 3.2**).

After a hidden batch effect analysis, two samples (Sko_13_1 and Kan_13_3) were found slightly deviated from other samples at 13 DAA from 'Kansas Queen' and 'Skofovged' varieties (**Supplementary Figure 3.4**). Because of the major dispersion of the data, we excluded both samples (Sko_13_1 and Kan_13_3) for further analyses.

An exploratory analysis and visualization of count data were performed to ensure high quality data with a previous pre-filtering of very low or no expressed genes (all genes with less than 10 reads in 3 samples were discarded). With an average of 20.07 ± 1.76 reads per sample, a total of 28,067 transcripts remained after filtering very low or no expressed genes from the initial 44,677 transcripts annotated in HFTH1 genome. Then, regularized-logarithm transformation (rlog) for negative binomial distribution was applied to the count matrix as variance stabilizing transformation (**Supplementary Figure 3.5**). From these analyses we can corroborate that most of the biological replicates clustered together showing little distance between them.

Differentially expressed genes between cultivars per developmental stage

The PCA analysis with the genes differentially expressed between cultivars at each time point clustered preferentially the samples collected at the same stage (**Supplementary Figure 6B**). The principal component 1 (PC1) captured 55 % of the variance, while the PC2 captured 18 %, which represents variance from the dataset. In the PCA the samples were displayed from early to late time point along the x-axis. Similarity between samples

were assessed by the construction of Euclidean distance matrix and represented as a heatmap and a Principal Component Analysis (PCA) plot (**Supplementary Figure 6A-B**).

DEGs were obtained by contrasting samples from different varieties at the same time of development, or at different development stages for the same variety. During the contrasts, the earlier point time at 13 and the spheroid or round shape variety 'Kansas Queen' were taken as reference.

When looking for DEGs along development, the major number of DEGs were found when the contrast compared the extreme stages (98 and 13 DAA), followed by the contrasts between 61 and 13 DAA (**Supplementary Table 3.3**).

Taking together all the DEG in the different contrasts, a total of 3,530 genes were differentially expressed in all the cultivars in 61 vs 13 DAA, while 1995 in 98 vs 61 DAA. When were analyzed the common genes differentially expressed between varieties at the same development stage, we found 131, 615 and 604 genes at 13, 61 and 98 DAA, respectively (**Figure 3.5**).

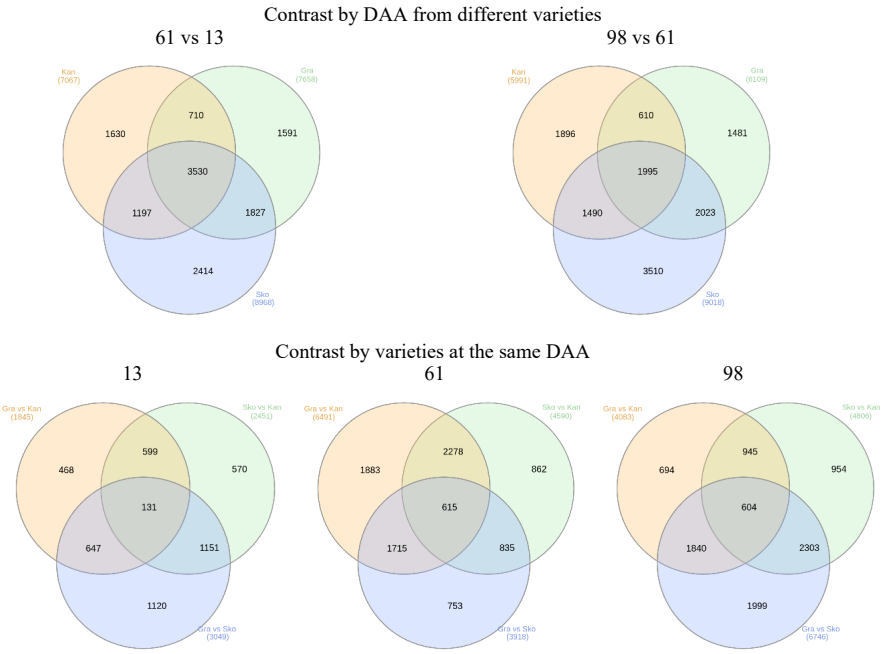


Figure 3.5. Venn diagramms from DEG contrast by varieties or DAA comparing samples pair-by-pair.

Differentially expressed genes along development

A time course-differential expression analysis (TC-DEA) considering cultivar and DAA as factors was performed to test for differences in gene expression along fruit development in each cultivar. A total of 12,551 transcripts were found differentially expressed.

In the TC-DEA results, genes with small *p* values from this test were those which at time points at 61 and 98 DAA showed a variety specific effect. Diagnostic graphs for the goodness of the treatment applied to the data are shown in **Supplementary Figure 3.7**.

Gene Annotation of differentially expressed genes

All the differentially expressed genes obtained in time course analysis with the packages were annotated using Gene Onthology (GO) terms, IPR, KEGG orthologs, KEGG pathways, non-redundant proteins in NCBI for all species and GDDH13 apple genome, TAIR *A. thaliana* orthologs and TREMBL (see in detail **Additional data 3.1**).

Correlation and R-squared values between TPMs with size, shape, and microscopy data along development.

Spearman's correlation coefficient was used to explain the relation between the transcripts obtained from the TC-DEA and the measurements provided for growth and parenchyma analysis.

Hundreds of genes among the 12,551 identified in time course-differential expression analysis (TC-DEA) had from moderate to strong correlation, considering values higher than 0.5 and lower than -0.5. When using the transcripts per million (TPM) and the values of the size and shape along fruit development (at points 13, 61 and 98 DAA), we found 2,550 genes with a positive correlation (r+) and 1,145 genes with a negative correlation (r-) with the FSI. Also, we found 1,992 genes (r+) and 3,925 (r-) genes with

TPM values correlated with fruit height values. Finally, the TPM in 1,797 genes (r+) and 3521 genes (r-) correlated with fruit width.

These genes were mostly involved in signaling pathways, cellular components, and biological processes, among them there were genes described for their role in fruit development such as hormone-related proteins: auxin response or induced factor (AUXs), ethylene response or induced factor (ETs), gibberellin (GAs) and Gamma aminobutyric acid (GABA). In addition, genes related to the regulation of leaf or fruit shape: *MADs-Box*, *YABBY* and *Ovate family protein* (*OFP*) genes (**Figure 3.6A** and **Supplementary Table 3.4**).

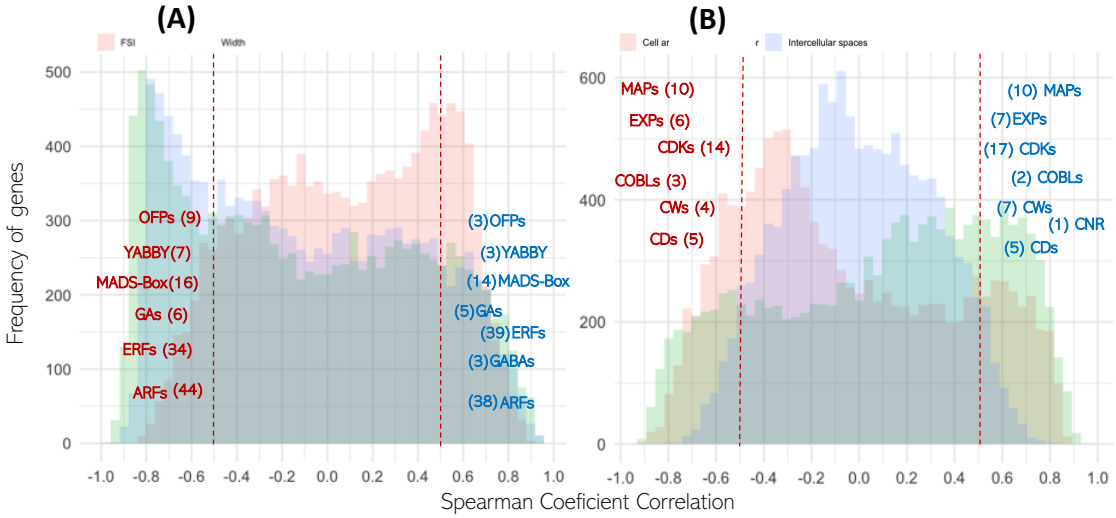


Figure 3.6. Histogram of Spearman correlation values derived from count matrix normalized in TPM (transcripts per million) and data from fruit measurements at points 13, 61 and 98 DAA. **(A)**, FSI, height and width correspond to macroscopic measurements at stages 13, 61 and 98. **(B)**, Cell area, cell number and intercellular spaces are measurements obtained by microscopic analysis of cells from parenchyma tissue at stages 61 and 98 DAA. The annotated genes are from TC-DEA (Time Differential Expression Analysis) results.

When looking at the correlation between gene TPM and cellular measurements (cell area, cell number and IS) at 61 and 98 DAA data, we identified 2060 genes (r+) and 2202 genes (r-) for cell area, and 3062 genes (r+) and 1964 genes (r-) for cell number. For the

IS we found 630 genes (r+) and 580 genes (r-). Some of these genes were related to cellular components, hormones, biosynthesis of organic compounds, cell division control (CDs), cell wall components (CWs), *cyclin-dependent Kinases* (*CDKs*), cell expansion (EXP) and *mitogen-activated proteins* (*MAPs*) (**Figure 3.6B** and **Supplementary Table 3.4**).

We used the coefficient of determination (R-squared or R^2), to determine the proportion of the total variance of variable explained by the regression model. We looked at the R^2 of all the annotated genes identified in the TC-DEA (12,551) for the first analysis, with especial attention to the ones also identified in the GWAS analysis (**Chapter 2**) and selected those with R^2 below -0.5 and above 0.5.

For the first analysis we considered the annotated genes from the TC-DEA and calculated the R^2 between TPMs and FSI, height and width. Sixty-four of the 12,551 annotated genes were related to hormones and regulators in floral organ, leaf, and fruit shape (**Supplementary Figure 3.8A** and **Table 3.5**). Twenty-nine genes corresponded to auxin proteins (AUXs), among them, the one with highest R^2 value (0.88) was the predicted gene *auxin-responsive protein IAA26-like*, with down-regulated activity. Eighteen genes were related to ethylene (ETs), the predicted *ethylene-responsive transcription factor ERF073-like* gene had the highest R^2 value (0.80) and is down-regulated. Two genes were annotated as gibberellin receptor (GAs) with an average R^2 of 0.7 and were downregulated. One gene *GABA* (Predicted: *Gamma aminobutyric acid transporter 1-like*), Nine were annotated as MADS-Box genes, being the *MADS-box transcription factor 14-like* gene the one with the highest R^2 (0.85) and down-regulated. Three genes of the *ovate protein family* (*OFPs*) had an average R^2 of 0.59 with up-regulated activity. Two

genes were annotated as axial regulator *YABBY 1-like* (*YABBYs*) and with a down-regulated activity.

R^2 values for between TPMs and microscopy parameters found 18 genes associated to cell proliferation, cell expansion and regulation of cell division (**Supplementary Figure 8B and Table 8**). Three of them (*CDPK-related kinase 7-like isoform, programmed cell death protein 4,* and *mitogen-activated protein kinase kinase kinase A-like*) with an average R^2 value (0.62) for cell area and 15 genes with average R^2 (0.57) for cell number measurement (**Supplementaries Figure 3.8** and **Table 3.6**).

For the second analysis we used the annotated genes from TC-DEA and from candidate genes obtained by GWAS, to calculate the R^2 between TPMs with FSI, height and width. A total of 31 genes related to fruit development and shape were filtered out (see **Supplementary Table 2.1-Chapter 2**). For example, 15 candidate genes were associated with quantitative size traits (such as area, maximum width, maximum height, and width-mid height), their average R^2 value was 0. 22 for height and 0.17 for width. The highest value (0.70 and 0.59) was for the gene HF02644 (PREDICTED: *auxin-induced protein 15A*), whose expression is down-regulated (**Figure 3.7** and **Supplementary Table 3.7**).

Seventeen candidate genes resulted associated for quantitative traits fruit shape (FSII, FSIINT, C, DFB, PAMa and CAT-own measurements), obtaining an average R^2 of 0. 18 with the FSI. Some of these genes were related to fruit shape regulation, such as the ovate protein family gene (HF43536: *OFP4-like*) with down-regulated activity. Also, the gene (HF37846: *transcription factor WRKY 33*) is involved in hormone regulation, with down-regulated activity. In addition, two of the genes identified were related to cell division (protein *DEK isoform X1* and *proliferating cell nuclear antigen (PCNA)*) with an

R^2 value ~0.63 and ~0.47 for height and width, respectively (**Figure 3.7A-C** and **Supplementary Table 3.7**).

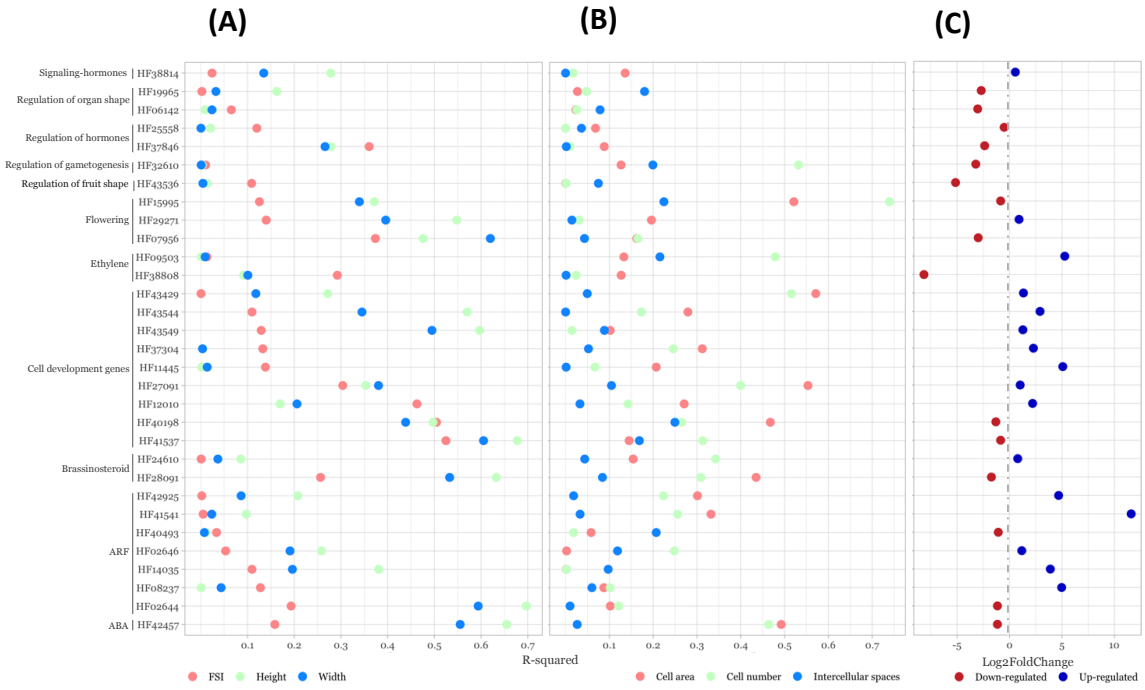


Figure 3.7. Dot plot of R-squared and Log2FoldChange values of genes selected by GWAS and TC-DEA results. **(A)** Corresponds to R-squared values obtained from analysis between count matrix in TPM and the FSI, height and width (development data) at the 13, 61 and 98 DAA data. **(B)** R-squared values obtained from analysis between count matrix in TPM and the cell area, cell number and intercellular spaces (parenchyma tissue analysis) of the 61 and 98 DAA data. **(C)** Log2foldchange values of genes selected, the gray line separates down-regulated and up-regulated genes. Acronyms, Auxin response factor (ARF), Acid Abcisic (ABA).

In the TPM vs parenchymal analysis measurements (cell area, cell number and IS), nine genes involved in cell development or proliferation were identified with an average R^2 of 0.32 for cell area, 0.23 for cell number, and for IS below 0.1. two of the 9 genes (HF43429, PREDICTED: *mitogen-activated protein kinase kinase kinase A-like*; and HF27091, PREDICTED: probable *serine/threonine-protein kinase WNK11*) were up-regulated and one (HF40198, PREDICTED: *proliferating cell nuclear antigen*) was down-regulated, all three with R^2 values higher than 0.47 for cell area. Three of the 31 genes

(HF15995, PREDICTED: probable *serine/threonine-protein kinase WNK9* isoform X1; HF32610, PREDICTED: *SNF1-related protein kinase regulatory subunit gamma-like PV42a;* and HF43429) had R^2 values higher than 0.51 for cell number (**Figure 3.7B-C** and **Supplementary Table 3.8**).

Comparative gene expression

The RNA-seq data was validated by qPCR of the HF43536 gene, obtaining a r-squared of 0.8591 between Ct and log2(TPM) values (**Supplementaries Table 2.6** and **Figure 2.5 from Chapter 2**).

For each of the genes with high R2 values described above, we tested for significant differences between pairs of cultivars in the TPM values at each data point (**Supplementary Table 3.9** and **Table 3.10**)

For example, at 13 DAA we found differences in the expression of HF38270 (Gid1C-like gibberellin receptor) and HF43429 (mitogen-activated protein kinase A-like) in KAN vs SKO genotypes. Similarly, the gene HF20655 (CDPK-related kinase 7-like isoform X2) was differentially expressed in the comparison GRA vs SKO. At point 61 DAA seven AUX genes (HF02639, HF02644, HF15394, HF22057, HF24542, HF34014 and HF42071) showed differential expression in GRA vs KAN and GRA vs SKO.

For the ethylene genes, seven genes (HF00548, HF09724, HF15130, HF27000, HF32335, HF39376 and HF41584) showed differences in GRA vs KAN (5 genes), GRA vs SKO (2 genes) and KAN vs SKO (4 genes). The gene HF38270, a gibberellin receptor, was differentially expressed in GRA vs SKO. Also, two MADS-box genes (HF24734 and HF34993) were differentially expressed in GRA vs KAN and KAN vs SKO. The gene HF28238, annotated as a transcription repressor *OFP12-like*, was differentially expressed in GRA vs KAN and GRA vs SKO (**Supplementary Figure 3.9** and **Table 3.9**).

At point 98 DAA, 10 hormone-related genes and shape regulators in the plant, were differentially expressed in GRA and KAN (HF02639, HF33235 and HF39430), GRA vs SKO (HF08647, HF33235, HF08568, HF21474, HF27287, HF15941, and HF34993) and between KAN and SKO (HF19076 and HF27287) (**Supplementary Figure 3.9** and **Table 3.9**).

Most of the candidate genes obtained from GWAS and TC-DEA showed differential expression in at least one of the DAA points and in contrasts between cultivars. Such differences were not observed in six of the genes (**Figure 3.8** and **Supplementary Table 3.10**).

If we describe the differential expression within cultivars, in the SKO genotype the genes with the highest level of expression were HF12010, HF11445, HF37304, HF43544, HF43429, HF09503, HF29271, HF32610, HF38814, while the genes with lowest gene expression were HF42457, HF02644, HF43536, HF25558. In GRA the highest expression was in HF40493, HF07956, HF43536, and HF06142 and the lowest in HF14035 and HF43544.

And finally, in KAN the highest expression in one of the three DAA points were (HF42457, HF02644, HF08237, HF14035, HF42925, HF28091, HF24610, HF43549, HF38808, HF37846, HF25558) and the lowest in (HF02646, HF40493, HF08237, HF29271, HF06142, HF19965, HF38814) (**Figure 3.8** and **Supplementary Table 3.10**).

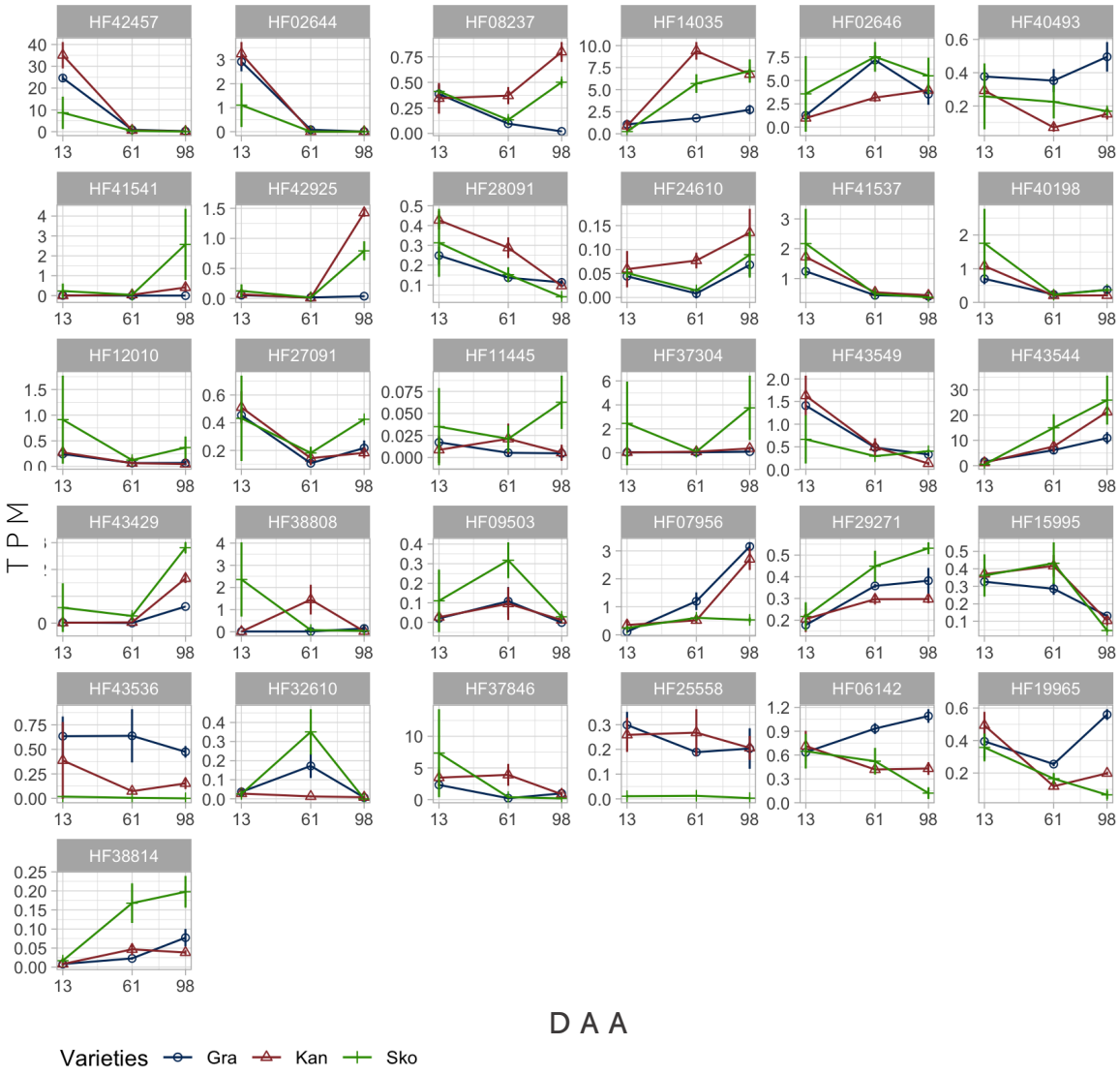


Figure 3.8. Lines plot of the comparative analysis of the count matrix normalized in TPM of the selected genes by the gene annotations obtained in the GWAS and TC-DEA. Comparing the expression at stages 13, 61 and 98 DAA of the three varieties studied Grand'mere (GRA), Kansas Queen (KAN) and Skovfoged (SKO).

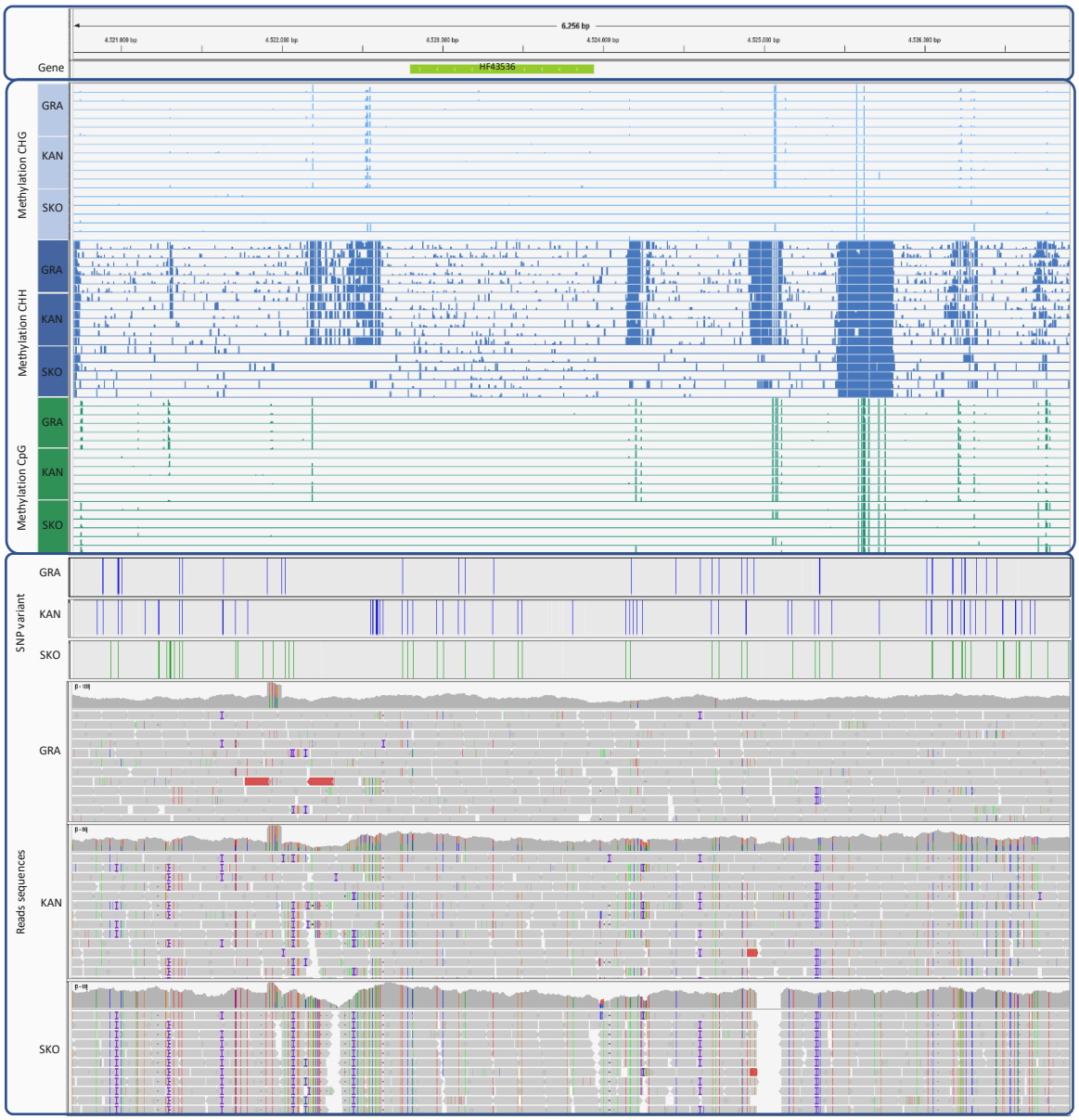


Figure 3.9. Analysis of the HF43536 (Ovate family gene) region, describing the methylation and SNP variant of the region. On top show the gene in left sense in a broad region of 6,256 pb in the chromosome 11: 4,520,702 - 4,526,958 of double haploid HFTH1 v1.0 genome. Below it shows different methylation levels in CHG, CHH and CpG from 6 samples per variety, following the SNP variant and the reads sequences per variety. Color references: for SNP variant blue as heterozygous variant, green as homozygous variant, and grey as homozygous reference genome. For reads sequences the color correspond to mutations related to reference.

Ovate family protein

Based on DEGs contrasts and the expression level of the HF43536 (PREDICTED: *OFP4-like transcription repressor*) gene, we have identified differences in the three genotypes. This gene was found as candidate for shape in the GWAS analysis in Chapter 2. This gene was not expressed in SKO, with oblong fruit shape, in any of the three stages evaluated, while was the one with the highest level of expression in the three data points in the oblate cultivar GRA, being below the spheroid genotype KAN (**Figure 3.8** and **Supplementary Table 3.10**).

Whole genome bisulfite data obtained from leaves of the three cultivars was analyzed to search for possible differentially methylated sites in this gene. We found that this gene was methylated in GRA and KAN from -222 to -291 nt of TSS, corresponding to CHH context) and at -260, -261, -294 and -295 nt of TSS in CpG context. These sites were not methylated in SKO. In contrast, 1 out of 6 samples at -261 nt of TSS was methylated in CpG context. These changes occurred in the promoter of the genes, which is 1kb upstream of TSS (**Figure 3.9**).

The whole-genome DNA sequence alignment reveal a possible structural variation in the promoter of the gene in SKO which needs to be validated (**Figure 3.9** and **Supplementary Table 3.11**).

DISCUSSION

Development of the apple fruit

Fruit growth and development occur from flower fertilization to fruit ripening. Here we have studied fruit development in three fruit shape typologies, linking morphology and parenchyma phenotypes with gene expression.

In **Chapter 1**, we used a random forest test to ultimately identify the FSI as the most relevant variable to assign fruits into categorical fruit shape classes. In this chapter we evaluated the FSI (fruit shape index) along fruit growth in nine cultivars with low, medium, and high FSI values obtained in **Chapter 1**. FSI trend was used to select the three cultivars with more contrasting FSI values (‘Grand'mere’, GRA (oblate or flat); ‘Kansas Queen’ (KAN) spheroid or round; and ‘Skovfoged’ (SKO) oblong) as well as the developmental stages for further analysis (0, 61 and 98 DAA).

The fruits of the pomoidae species are formed by two distinct parts: the core corresponding to the expansion of the ovary (which is homologous to other fruits as, for example, the tomato) and the cortex (hypanthium) or edible portion of the fruit which is derived from the fused base of stamens, petals, and sepals (Janssen et al., 2008). Along fruit development, both parts expand due first to cell division and later to cell expansion to reach the final size and shape. However, shape differences were already observed in these three cultivars in the ovary (0 DAA). GRA had larger width and already showed a flattened appearance in height, in KAN the two dimensions were similar, and SKO showed an elongation in the apical zone.

In addition to the morphology traits acquired in **Chapter 1**, we added parenchyma microscopy observations (cell number, area, and intercellular spaces) as additional phenotypes. Some morphology and histology measures showed correlation, as is the case of fruit height and cell area, which showed positive correlation. By contrary, cell area correlated negatively with cell number.

Differential Expression Analysis

Total RNA obtained from the parenchyma of fruits at 13, 61 and 98 DAA was sequenced to analyze differences in gene expression linked to morphology or parenchyma cells data, what allowed the study of the variation of the transcripts along fruit development. Preliminary exploratory analyses and the results of the DEGs suggest that in the earlier stages, the samples are more similar than in the later stages, when the fruit is fully developed and has been exposed to externa factors longer. The closer relationships found in the Euclidean distance matrix and the lower number of DEGs found at the earlier stages may be because the differentiation process has only begun, but, in these younger organs, small changes will trigger the higher number of DEGs found at the later stage of 98 DAA, when the fruit organ is almost fully developed and can give place to the phenological divergences in size and shape found between varieties (Love et al., 2014). The DEGs analyzed by DAA point and variety contrasts identified genes common between contrasts, and which are specific. In this work we focused on the genes putatively related to fruit development and fruit shape based on their annotation.

Time course Differential expression analysis

Among the TC-DEA annotated genes with higher R^2 (i.e. with higher variance of the dependent variable (the phenotype) explained by the independent variable (the TPM) we found multiple phytohormones (R^2> 50 %): auxin response or induced factor (AUXs), ethylene response or induced factor (ETs), gibberellin (GAs) and Gamma aminobutyric acid (GABA). AUX are known to be involved in the regulation of various aspects of fruit development such as cell proliferation, cell expansion and fruit ripening (Srivastava and Handa 2005). In one study, auxins were determined as responsible for the final size of the apple (Devoghalaere et al., 2012). We found the ARF9 gene (HF08647, PREDICTED:

auxin response factor 9) associated with fruit height along growth and with down-regulated activity. According to de Jong et al., (2015) overexpression of the ARF9 gene in tomato reduced fruit size and had a down-regulated activity on cell production during early fruit development. Other relevant genes identified during fruit growth were those of the IAA group (*indole-3-acetic acid*), known as free auxins; we identified five 5 of them differentially expressed during growth. In Devoghalaere et al., (2012) increased the cortex or hypanthium zone of the fruit.

Ethylene is known to be responsible for the ripening of several fruit species such as melon, tomato, apple (Pereira et al., 2020; Liu et al., 2016; Yue et al., 2020). We identified fifteen ethylene-related proteins, one of them is the gene (HF13168, Predicted: *AP2-like ethylene-responsive transcription factor ANT*) with down-regulated activity. This gene is involved in the control of primary and secondary metabolism in growth and development, as well as in responses to environmental stimulation (Licausi et al., 2013).

The phytohormone gibberellin (GA) was also identified in genes differentially expressed during growth with R^2>0.53 in relation to fruit height and width, such as the gene HF15941 (*gibberellin receptor Gid1C-like*), described as nuclear GA insensitive dwarf1s (*GID1s*) receptors responsible for triggering degradation of *DELLAs* repressors. In Arabidopsis, at the early stage of fruit development they are transcriptionally active and play an important role in seed development and pod elongation (Gallego-Giraldo et al., 2014). In apple, GA applications at the fruit set stage induce fruit shape changes, showing a greater growth in both height and width (Nakagawa et al., 1968).

In addition to the hormones already mentioned, other hormones with lower R^2 values were also identified, such as jasmonates related to the fruit ripening process (Li et al.,

2017), brassinosteroids promoting cell proliferation, fruit ripening and senescence (Clouse, 2011), and abscisic acid (ABA) involved in fruit set abscission, but also found an endogenous concentration of ABA in the fruit cortex (Eccher et al., 2013).

Another group of genes linked to development and growth are the *MADs-Box* protein family; we identified nine of this group in the TC-DEA. In apple, they were characterized and classified within the *APETALA1* (*AP1*) and *AGAMOUS* (*AG*) groups, which show differential expression in the core, cortex and skin in young fruits (Yao et al. 1999).

When using the cell traits as phenotypes, the genes with R^2>0.5 identified were related to cell division and expansion, such as the *MAPKKK* (mitogen-activated protein kinase kinase kinase) gene cluster, whose function is the transduction of environmental and developmental signals, in addition to cell cycle progression (Jagodzik et al., 2018).

One of them is *NPK1*, which explains in a large proportion cell area and has up-regulated activity. According to Nishihama et al., (2001) it has activity in the M phase of cell division, which is essential for the formation of the cell plate and its lateral growth, and therefore is required for cytokinesis. In relation to cell expansion, 9 EXPANSINs (EXP) genes have been identified. These proteins are known to have a loosening activity, cell expansion and cell wall modification (Sampredro and Cosgrove, 2005). One of them is *MdEXPA20* associated with cell number down-regulated activity in the TC-DEA. In a study in apple, Zhang et al., (2014) found that *MdEXPA20* expression plays an important role in fruit development in relation to cell expansion during growth.

Comparison of gene expression in three fruit shapes

We found that some of the candidate genes identified in the GWAS analysis (**Chapter 2**) showed differential expression along development (**Figure 3.10**).

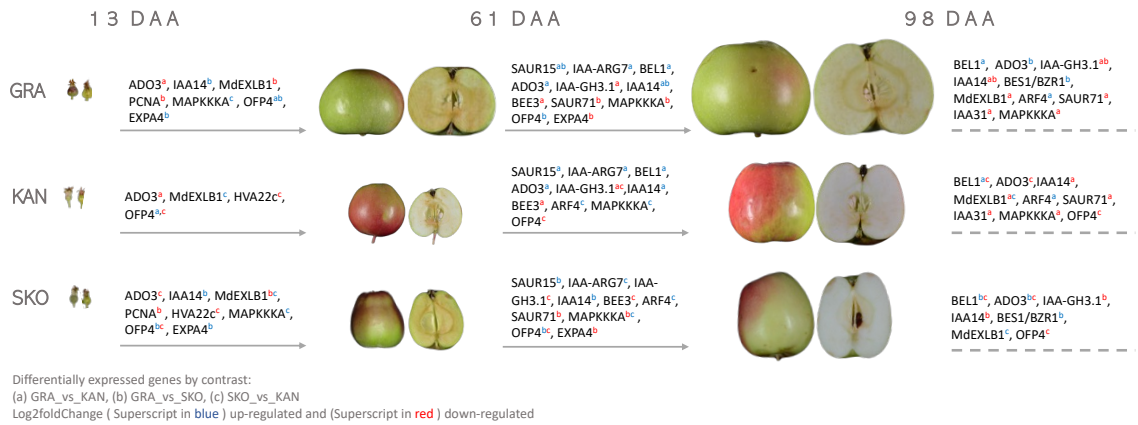


Figure 3.10. Schematic summary of candidate genes analyzed and filtered by GWAS and DEGs results, along the fruit development in three points (13, 61 and 98) DAA.

At an early developmental stage such as the 13 DAA point, an exponential level of cell proliferation is found in the fruit and in parallel the cell expansion genes initiate their transcription (Jansen et al., 2008). Here we found in the DEG contrast of GRA_vs_SKO, the *PCNA* gene, with function in cell proliferation, is differentially expressed and the number of normalized transcripts (TPM) in SKO has a level of expression above the other genotypes. The *MdEXLB1* and *EXPA4* genes described for their cell spreading function are also differentially expressed according to the GRA_vs_SKO and SKO_vs_KAN contrasts, with the SKO genotype showing a higher level of expression. According to the results of the cell measurements at the 0 DAA point, SKO genotype showed the highest cell area and the lowest number of cells. We do not have RNA data for the 0 DAA, but the analysis of the samples collected at 13 DAA (stage in which the fruit is still growing, as shown by the fruit morphology observation along development) identified genes with up-regulated activity linked to proliferation and regulation of cell division. The *IAA14* gene (auxin response protein) was associated with height and width and was differentially expressed between GRA (wider) and SKO (taller).

At 61 DAA the fruit is approximately in the middle of its development and fruit shape differences are evident (**Figure 3.2**). Auxin hormone proteins are still differentially expressed up-regulated (**SAUR15, IAA-ARG7, IAA14**) and down-regulated (**SAUR71** and **IAA-GH3.1**). The expression level for KAN in the genes (**IAA-GH3.1** and **IAA14**) and for GRA (**IAA-ARG7** and **ARF4**) are above the other two genotypes. Notably, the expression level of the **SAUR15** gene, explains 64% of fruit height and width by R-squared. This gene is in the candidate region of a QTN for maximum height (see **Chapter 2**, Supplementary Table S7).

The BEL1 gene, homologous to MdH1 in apple, was found in flowers, expanding leaves and expanding fruit, and experimental assays in transgenic Arabidopsis plants showed dwarfism, reduced fertility and changes in carpel and fruit shape (Dong et al., 2000). In our study, this gene is differentially expressed downregulated, showing an increase in GRA over the other genotypes. Furthermore, this gene is located in the candidate region of the QTN for maximum width (see **Chapter 2, Supplementary Table 2.4**).

As for genes controlling cell proliferation and expansion, we identified EXP4 gene (expansinA4-like precursor) up-regulated in the DEG contrast in GRA_vs_SKO with an R^2 =0.55 related to height along growth. This gene is in the QTN of FSIINT and CAT-own (related to fruit shape) (**Figure 10** and **Chapter 2, Supplementary Table 2.1**).

At 98 DAA, the differential expression of the above-mentioned genes identified between the different contrasts is similar to those at 61DAA. Fruit growth decreases the expression of cell expansion genes, and the ripening stage triggers different biosynthesis processes, such as ethylene, auxin and conversion of starch to fructose (Janssen et al.,2008; Bussatto et al. 2017).

During fruit development, we also identified genes that differentially expressed in all the DAA points studied, such as ADO3, MAPKKKK and OFP that have a high relationship (**Supplementary Table 3.5** and **3.6**) with fruit height, width and FSI, and at the cellular level with cell area and cell count (**Figure 3.7** and **3.10**).

Candidate gene for fruit shape in apple

The *OFP4* (ovate family protein 4) gene, a protein belonging to the family of transcription factors found only in plants, was identified in all differential gene expression analyses as well as in the GWAS results. In dicots, *OFP* family genes control fruit shape and secondary cell wall biosynthesis (Schmitz et al., 2015). In addition, their molecular function acts as a transcriptional repressor, i.e. an elevated expression level can suppress the activity of other genes, resulting in a change in fruit shape (Wang et al. 2016). According to DEGs contrasts, the GRA genotype (oblate) has this gene has a higher expression in th three developmental stages compared to KAN and SKO. This gene is candidate for the FSIINT and CAT-own QTNs which describe fruit shape and is located within a 9.7 kb haploblock on chromosome 11. In addition, based on the observed phenotypes, two SNP molecular markers could determine fruit type in at least two shapes (round and flat) (see **Chapter 2, Figure 2.3,** and **Supplementary Table 2.2**).

In other fruit species, such as peach, a candidate gene (*PpOFP1*) has been identified, which transcriptional activity can repress vertical elongation in flat fruits at the early stage of development (Zhou et al., 2020). In melon, the *CmFSI8/CmOFP13* locus encoding the *OVATE* protein orthologous to *AtOFP1* has been shown to be responsible for fruit shape by ectopic overexpression in Arabidopsis resulting in leaf shape changes with a kidney-like appearance or shortened siliques (Ma et al., 2021). In tomato, three

loci regulate ovary and fruit elongation at different stages, the interaction between *SUN, OVATE* or *fs8.1* has a direct effect on fruit shape (Wu et al. 2015).

The alignment of the whole genome DNA sequence of the three cultivars revealed a deletion in the promoter of the *OFP4*. If this polymorphism is validated and the association is confirmed in a larger dataset, a molecular marker to easily detect the polymorphism could be used for the early (positive or negative) selection of genotypes bearing flat fruits in breeding programs.

These results provide further insight into fruit shape in apple, studying within fruit development and in different shape types. Although there are differences at the cellular and gene expression levels, validation is needed to confirm the functional role of each candidate gene.

CONCLUSION

The data presented in this chapter revealed differences in three fruit shape typologies in apple along fruit development. Candidate genes involved in physiological processes with role in fruit growth have been identified, such as the auxin genes (*SAUR15, IAA-ARG7, IAA14, SAUR71 and IAA-GH3.1*), *EXP4* gene (cell expansion) and specific genes such as *BEL1* and the fruit shape regulator *OFP4*. The last one appeared to be involved in the fruit shape natural variation in apple fruit shape, as reflected by the correlation between expression level and the phenotypes studies (morphology and cell traits in the parenchyma). Future validation will determine the specific function of these genes.

REFERENCES

1. Inglis, P. W., Pappas, M. D. C. R., Resende, L. V., & Grattapaglia, D. (2018). Fast and inexpensive protocols for consistent extraction of high quality DNA and RNA from challenging plant and fungal samples for high-throughput SNP genotyping and sequencing applications. *PloS one*, *13*(10), e0206085.
2. Inzé, D., & De Veylder, L. (2006). Cell cycle regulation in plant development. *Annu. Rev. Genet.*, *40*, 77-105.
3. Jagodzik, P., Tajdel-Zielinska, M., Ciesla, A., Marczak, M., & Ludwikow, A. (2018). Mitogen-activated protein kinase cascades in plant hormone signaling. *Frontiers in plant science*, *9*, 1387.
4. Janssen, B. J., Thodey, K., Schaffer, R. J., Alba, R., Balakrishnan, L., Bishop, R., ... & Ward, S. (2008). Global gene expression analysis of apple fruit development from the floral bud to ripe fruit. *BMC Plant Biology*, *8*(1), 1-29.
5. Jung, M., Keller, B., Roth, M., Aranzana, M. J., Auwerkerken, A., Guerra, W., ... & Patocchi, A. (2022). Genetic architecture and genomic predictive ability of apple quantitative traits across environments. *Horticulture research*, *9*.
6. Kanehisa, M., Sato, Y., Kawashima, M., Furumichi, M., & Tanabe, M. (2016). KEGG as a reference resource for gene and protein annotation. *Nucleic acids research*, *44*(D1), D457-D462.
7. Kim, D., Langmead, B., & Salzberg, S. L. (2015). HISAT: a fast spliced aligner with low memory requirements. *Nature methods*, *12*(4), 357-360.
8. Kotoda, N., Wada, M., Komori, S., Kidou, S. I., Abe, K., Masuda, T., & Soejima, J. (2000). Expression pattern of homologues of floral meristem identity genes LFY and AP1 during flower development in apple. *Journal of the American Society for Horticultural Science*, *125*(4), 398-403.
9. Krueger, F. (2015). Trim galore. *A wrapper tool around Cutadapt and FastQC to consistently apply quality and adapter trimming to FastQ files*, *516*(517).
10. Krueger, F., & Andrews, S. R. (2011). Bismark: a flexible aligner and methylation caller for Bisulfite-Seq applications. *bioinformatics*, *27*(11), 1571-1572.

11. Lamesch, P., Berardini, T. Z., Li, D., Swarbreck, D., Wilks, C., Sasidharan, R., ... & Huala, E. (2012). The Arabidopsis Information Resource (TAIR): improved gene annotation and new tools. Nucleic acids research, 40(D1), D1202-D1210.
12. Langmead, B., & Salzberg, S. L. (2012). Fast gapped-read alignment with Bowtie 2. *Nature methods*, *9*(4), 357-359.
13. Lassmann, T., Hayashizaki, Y., & Daub, C. O. (2011). SAMStat: monitoring biases in next generation sequencing data. *Bioinformatics*, *27*(1), 130-131.
14. Leek, J. T., Johnson, W. E., Parker, H. S., Jaffe, A. E., & Storey, J. D. (2012). The sva package for removing batch effects and other unwanted variation in high-throughput experiments. *Bioinformatics*, *28*(6), 882-883.
15. Li H, Ruan J, Durbin R. (2008) Mapping short DNA sequencing reads and calling variants using mapping quality scores. *Genome Research* **18**:1851-8.
16. Li, H., & Durbin, R. (2009). Fast and accurate short read alignment with Burrows–Wheeler transform. *bioinformatics*, *25*(14), 1754-1760.
17. Li, H., & Durbin, R. (2010). Fast and accurate long-read alignment with Burrows–Wheeler transform. *Bioinformatics*, *26*(5), 589-595.
18. Li, T., Xu, Y., Zhang, L., Ji, Y., Tan, D., Yuan, H., & Wang, A. (2017). The jasmonate-activated transcription factor MdMYC2 regulates ETHYLENE RESPONSE FACTOR and ethylene biosynthetic genes to promote ethylene biosynthesis during apple fruit ripening. *The Plant Cell*, *29*(6), 1316-1334.
19. Liao, Y., Smyth, G. K., & Shi, W. (2014). featureCounts: an efficient general purpose program for assigning sequence reads to genomic features. *Bioinformatics*, *30*(7), 923-930.
20. Licausi, F., Ohme-Takagi, M., & Perata, P. (2013). APETALA 2/Ethylene Responsive Factor (AP 2/ERF) transcription factors: Mediators of stress responses and developmental programs. *New Phytologist*, *199*(3), 639-649.
21. Liu, M., Gomes, B. L., Mila, I., Purgatto, E., Peres, L. E., Frasse, P., ... & Pirrello, J. (2016). Comprehensive profiling of ethylene response factor expression identifies ripening-associated ERF genes and their link to key regulators of fruit ripening in tomato. *Plant Physiology*, *170*(3), 1732-1744.
22. Love, M. I., Huber, W., & Anders, S. (2014). Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome biology*, *15*(12), 1-21.

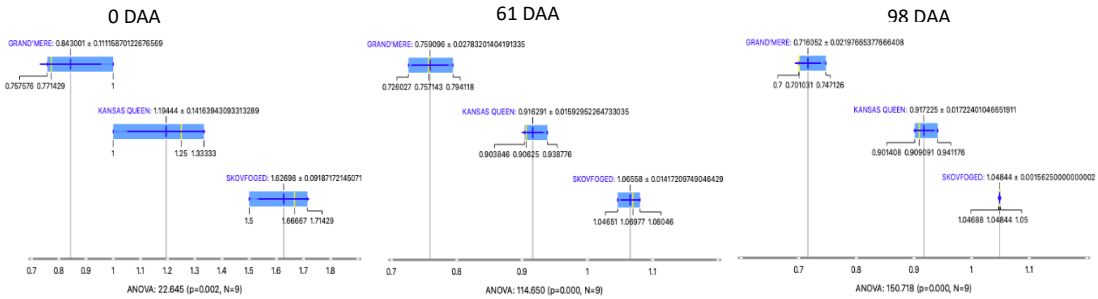
23. Ma, H., & DePamphilis, C. (2000). The ABCs of floral evolution. *Cell*, *101*(1), 5-8.
24. Ma, J., Li, C., Zong, M., Qiu, Y., Liu, Y., Huang, Y., ... & Wang, J. (2022). CmFSI8/CmOFP13 encoding an OVATE family protein controls fruit shape in melon. *Journal of Experimental Botany*, *73*(5), 1370-1384.
25. Malladi, A., & Johnson, L. K. (2011). Expression profiling of cell cycle genes reveals key facilitators of cell production during carpel development, fruit set, and fruit growth in apple (Malus× domestica Borkh.). *Journal of experimental botany*, *62*(1), 205-219.
26. Mauxion, J. P., Chevalier, C., & Gonzalez, N. (2021). Complex cellular and molecular events determining fruit size. *Trends in Plant Science*, *26*(10), 1023-1038.
27. Nakagawa, S., Bukovac, M. J., Hirata, N., & Kurooka, H. (1968). Morphological studies of gibberellin-induced parthenocarpic and asymmetric growth in apple and Japanese pear fruits. *Journal of the Japanese Society for Horticultural Science*, *37*(1), 9-19.
28. Nakagawa, S., Bukovac, M. J., Hirata, N., & Kurooka, H. (1968). Morphological studies of gibberellin-induced parthenocarpic and asymmetric growth in apple and Japanese pear fruits. *Journal of the Japanese Society for Horticultural Science*, *37*(1), 9-19.
29. Nishihama, R., Ishikawa, M., Araki, S., Soyano, T., Asada, T., & Machida, Y. (2001). The NPK1 mitogen-activated protein kinase kinase kinase is a regulator of cell-plate formation in plant cytokinesis. *Genes & Development*, *15*(3), 352-363.
30. Patil, I. (2021). Visualizations with statistical details: The'ggstatsplot'approach. *Journal of Open Source Software*, *6*(61), 3167.
31. Pereira, L., Santo Domingo, M., Ruggieri, V., Argyris, J., Phillips, M. A., Zhao, G., ... & Garcia-Mas, J. (2020). Genetic dissection of climacteric fruit ripening in a melon population segregating for ripening behavior. *Horticulture research*, *7*.
32. Pruitt, K. D., Tatusova, T., & Maglott, D. R. (2007). NCBI reference sequences (RefSeq): a curated non-redundant sequence database of genomes, transcripts and proteins. *Nucleic acids research*, *35*(suppl_1), D61-D65.
33. Sampedro, J., & Cosgrove, D. J. (2005). The expansin superfamily. *Genome biology*, *6*(12), 1-11.

34. Schindelin, J., Arganda-Carreras, I., Frise, E., Kaynig, V., Longair, M., Pietzsch, T., ... & Cardona, A. (2012). Fiji: an open-source platform for biological-image analysis. *Nature methods*, *9*(7), 676-682.
35. Schmitz, A. J., Begcy, K., Sarath, G., & Walia, H. (2015). Rice Ovate Family Protein 2 (OFP2) alters hormonal homeostasis and vasculature development. *Plant Science*, *241*, 177-188.
36. Serrani, J. C., Fos, M., Atarés, A., & García-Martínez, J. L. (2007). Effect of gibberellin and auxin on parthenocarpic fruit growth induction in the cv Micro-Tom of tomato. *Journal of Plant Growth Regulation*, *26*(3), 211-221.
37. Srivastava, A., & Handa, A. K. (2005). Hormonal regulation of tomato fruit development: a molecular perspective. *Journal of plant growth regulation*, *24*(2), 67-82.
38. Team, R. C. (2020). *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing.
39. Thorvaldsdóttir, H., Robinson, J. T., & Mesirov, J. P. (2013). Integrative Genomics Viewer (IGV): high-performance genomics data visualization and exploration. *Briefings in bioinformatics*, *14*(2), 178-192.
40. UniProt Consortium. (2019). UniProt: a worldwide hub of protein knowledge. *Nucleic acids research*, *47*(D1), D506-D515.
41. Van der Auwera, G. A., Carneiro, M. O., Hartl, C., Poplin, R., Del Angel, G., Levy-Moonshine, A., ... & DePristo, M. A. (2013). From FastQ data to high-confidence variant calls: the genome analysis toolkit best practices pipeline. *Current protocols in bioinformatics*, *43*(1), 11-10.
42. Wang, S., Chang, Y., & Ellis, B. (2016). Overview of OVATE FAMILY PROTEINS, a novel class of plant-specific growth regulators. *Frontiers in plant science*, *7*, 417.
43. Wang, Y., Clevenger, J. P., Illa-Berenguer, E., Meulia, T., van der Knaap, E., & Sun, L. (2019). A comparison of sun, ovate, fs8. 1 and auxin application on tomato fruit shape and gene expression. *Plant and Cell Physiology*, *60*(5), 1067-1081.
44. Wickham, H. (2016). Package 'ggplot2': elegant graphics for data analysis. *Springer-Verlag New York. doi*, *10*, 978-0.

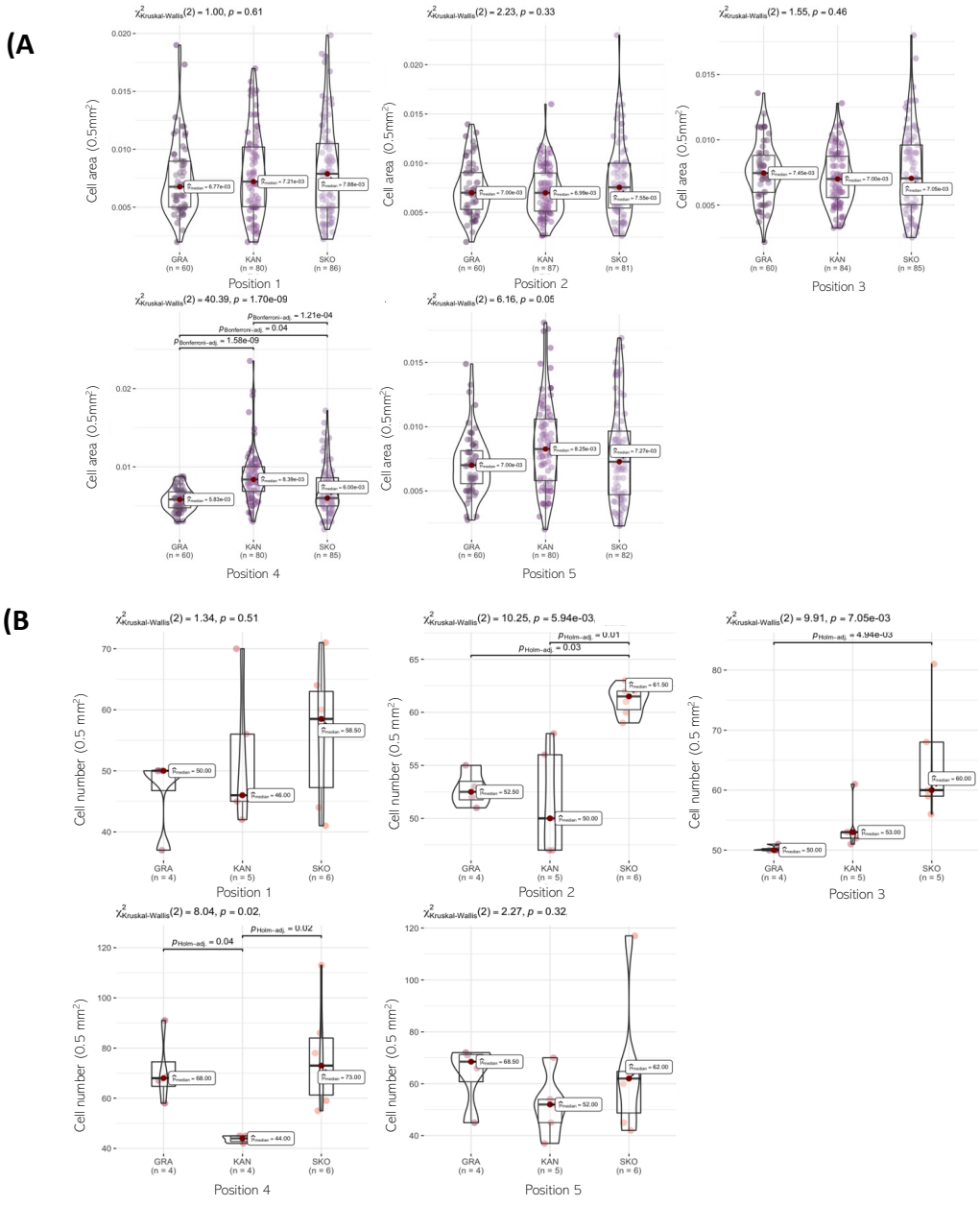
45. Wu, H., Xu, T., Feng, H., Chen, L., Li, B., Yao, B., ... & Conneely, K. N. (2015). Detection of differentially methylated regions from whole-genome bisulfite sequencing data without replicates. *Nucleic acids research*, *43*(21), e141-e141.
46. Wu, S., Clevenger, J. P., Sun, L., Visa, S., Kamiya, Y., Jikumaru, Y., ... & van der Knaap, E. (2015). The control of tomato fruit elongation orchestrated by sun, ovate and fs8. 1 in a wild relative of tomato. *Plant Science*, *238*, 95-104.
47. Wu, S., Clevenger, J. P., Sun, L., Visa, S., Kamiya, Y., Jikumaru, Y., ... & van der Knaap, E. (2015). The control of tomato fruit elongation orchestrated by sun, ovate and fs8. 1 in a wild relative of tomato. *Plant Science*, *238*, 95-104.
48. Yao, J. L., Dong, Y. H., & Morris, B. A. (2001). Parthenocarpic apple fruit production conferred by transposon insertion mutations in a MADS-box transcription factor. *Proceedings of the National Academy of Sciences*, *98*(3), 1306-1311.
49. Yao, J. L., Dong, Y. H., Kvarnheden, A., & Morris, B. (1999). Seven MADS-box genes in apple are expressed in different parts of the fruit. *Journal of the American Society for Horticultural Science*, *124*(1), 8-13.
50. Yao, J. L., Xu, J., Cornille, A., Tomes, S., Karunairetnam, S., Luo, Z., ... & Gleave, A. P. (2015). A micro RNA allele that emerged prior to apple domestication may underlie fruit size evolution. *The Plant Journal*, *84*(2), 417-427.
51. Yao, J. L., Xu, J., Tomes, S., Cui, W., Luo, Z., Deng, C., ... & Gleave, A. P. (2018). Ectopic expression of the PISTILLATA homologous MdPI inhibits fruit tissue growth and changes fruit shape in apple. Plant Direct 2018: 1–11.
52. Yue, P., Lu, Q., Liu, Z., Lv, T., Li, X., Bu, H., ... & Wang, A. (2020). Auxin-activated MdARF5 induces the expression of ethylene biosynthetic genes to initiate apple fruit ripening. *New Phytologist*, *226*(6), 1781-1795.
53. Zhang, L., Hu, J., Han, X., Li, J., Gao, Y., Richards, C. M., ... & Cong, P. (2019). A high-quality apple genome assembly reveals the association of a retrotransposon and red fruit colour. *Nature communications*, *10*(1), 1-13.
54. Zhang, S., Xu, R., Gao, Z., Chen, C., Jiang, Z., & Shu, H. (2014). A genome-wide analysis of the expansin genes in Malus× Domestica. *Molecular genetics and genomics*, *289*(2), 225-236.

55. Zhou, H., Ma, R., Gao, L., Zhang, J., Zhang, A., Zhang, X., ... & Han, Y. (2021). A 1.7-Mb chromosomal inversion downstream of a PpOFP1 gene is responsible for flat fruit shape in peach. *Plant biotechnology journal*, *19*(1), 192-205.

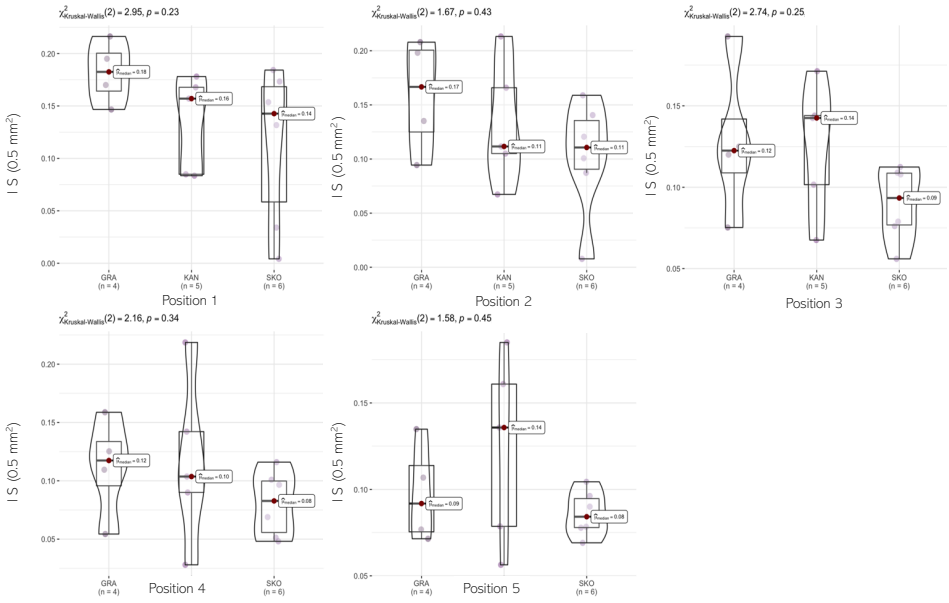
SUPPLEMENTARY MATERIAL



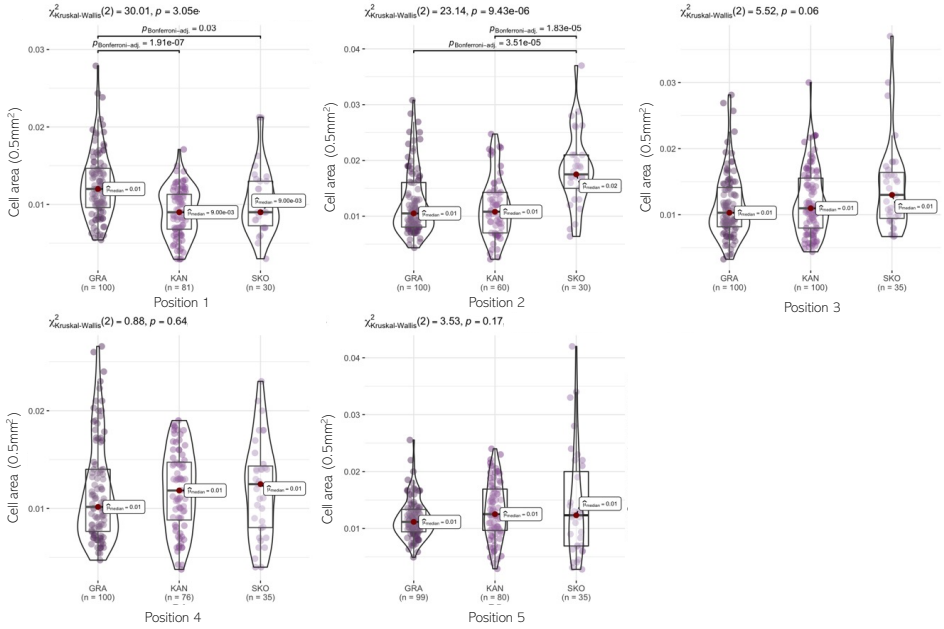
Supplementary Figure 3.1. Horizontal bar plot of the FSI values at the points 0, 61 and 98 DAA in three genotypes, each one represents a shape fruit, as 'Grand'mere' (oblate), 'Kansas Queen' (spheroid) and 'Skovfoged' (oblong).



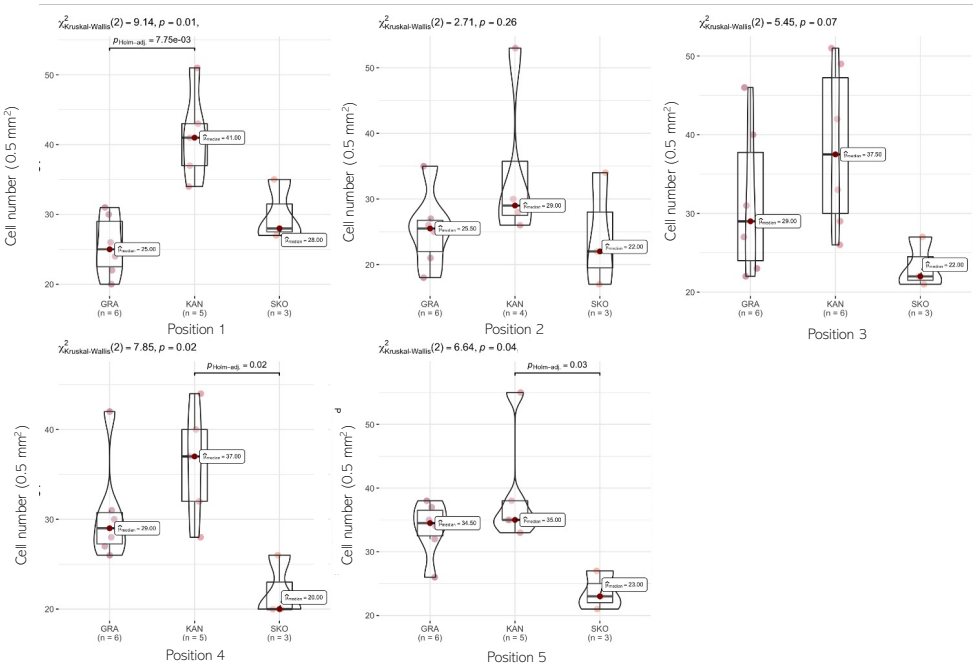
(C)


Position 1
Position 2
Position 3
Position 4
Position 5
GRA
KAN
SKO

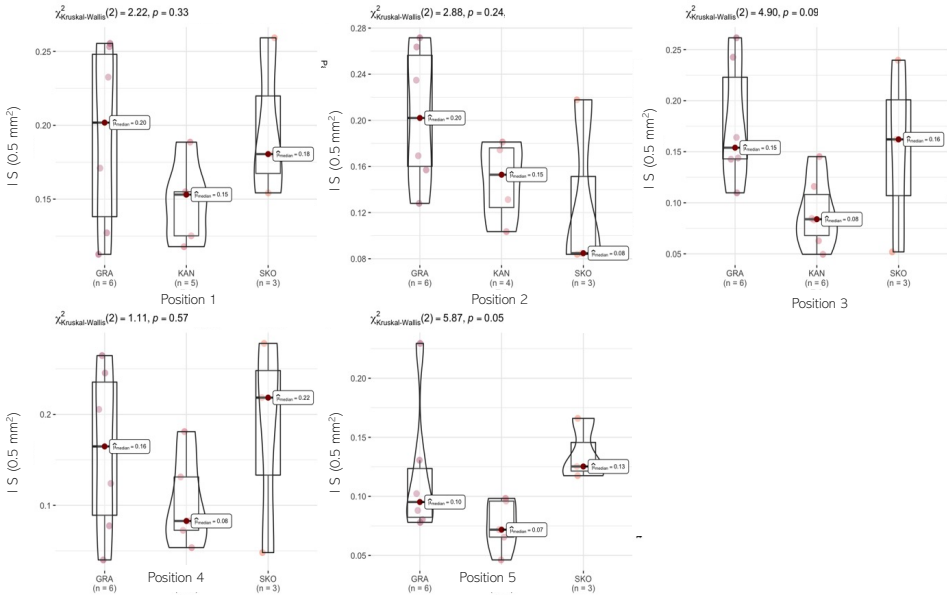
(D)


Position 1
Position 2
Position 3
Position 4
Position 5
GRA
KAN
SKO

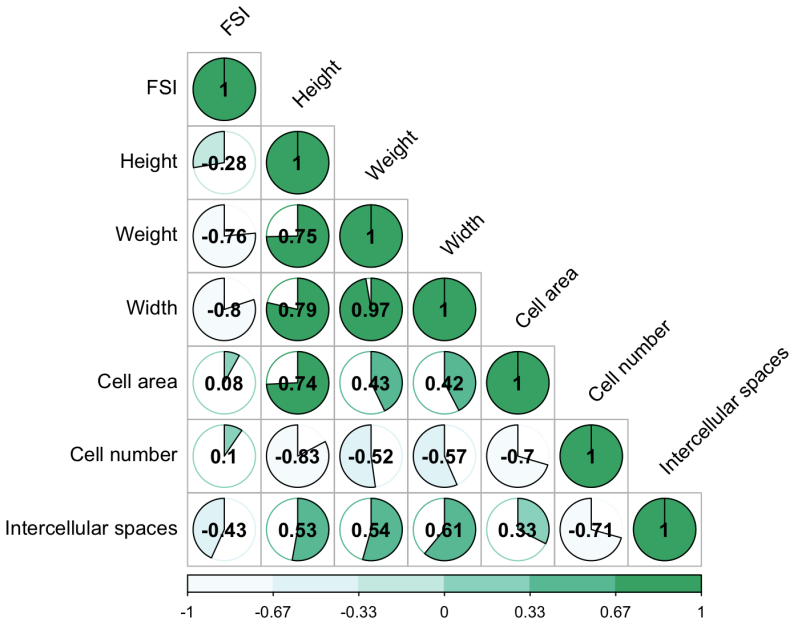
(E)


Position 1
Position 2
Position 3
Position 4
Position 5
GRA
KAN
SKO

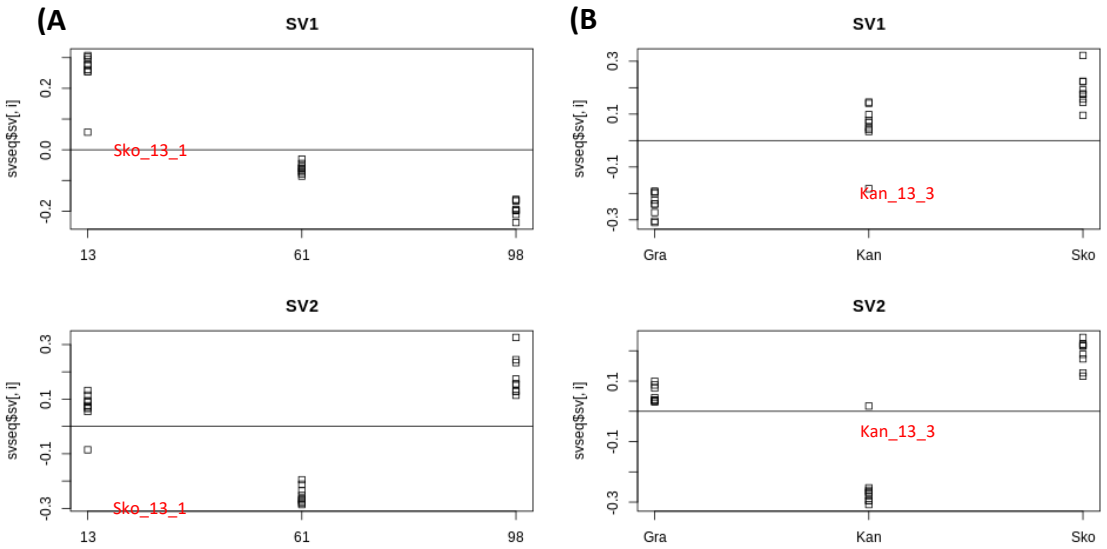
(F)



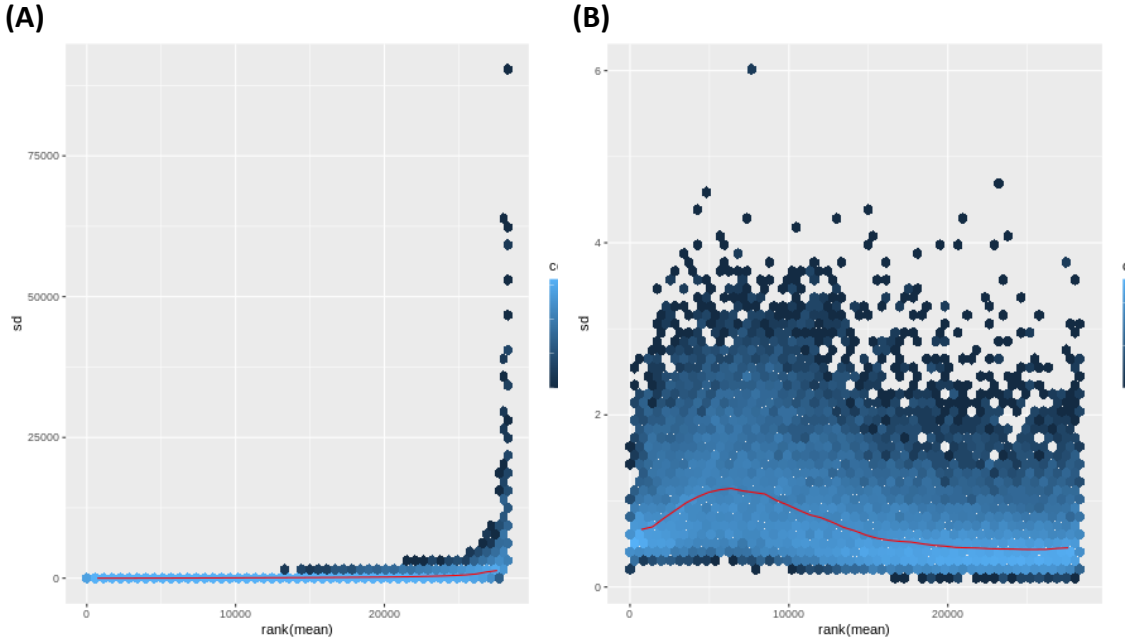
Supplementary Figure 3.2. Boxplot of parenchyma tissue analysis for 61 and 98 DAA at 5 positions along the hypanthium area by longitudinal sections of three apple varieties, Grand'mere (GRA), Kansas Queen (KAN) and Skovfoged (SKO) taking measures such as cell area (0.5 mm^2), cell number (0.5 mm^2) and intercellular spaces (0.5 mm^2). 61 DAA corresponds to plots (A), (B) and (C) and for 98 DAA, (D), (E) and (F).



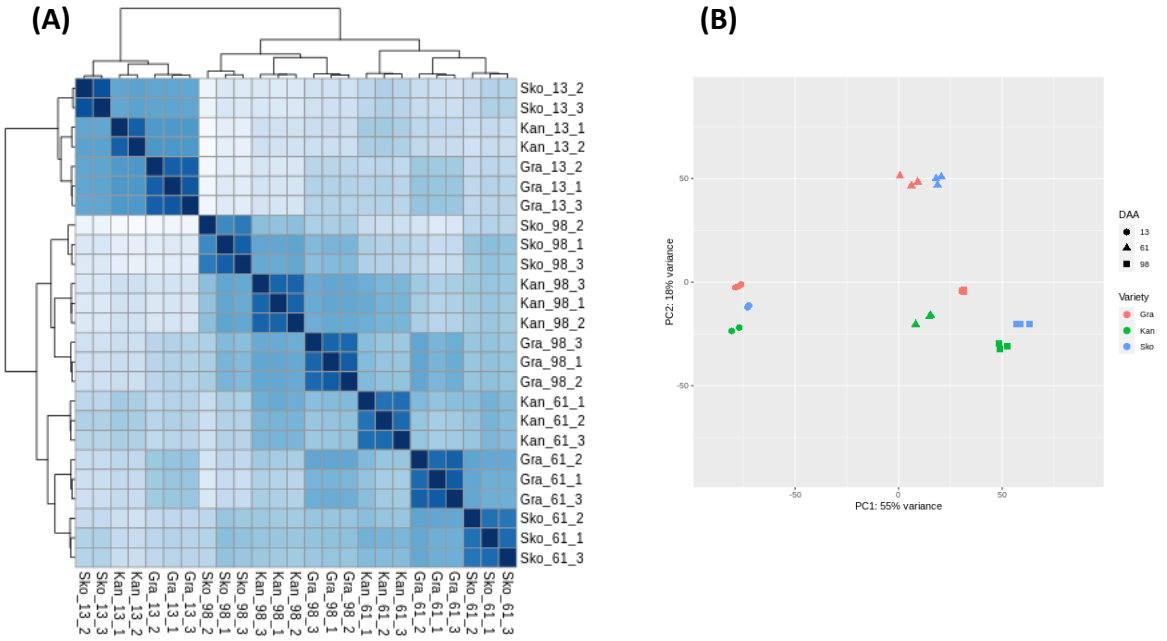
Supplementary Figure 3.3. Correlation plot of the variables taken from development measurements (FSI, height, weight, and width) and the microscopy analysis (cell area, cell number, and intercellular spaces).



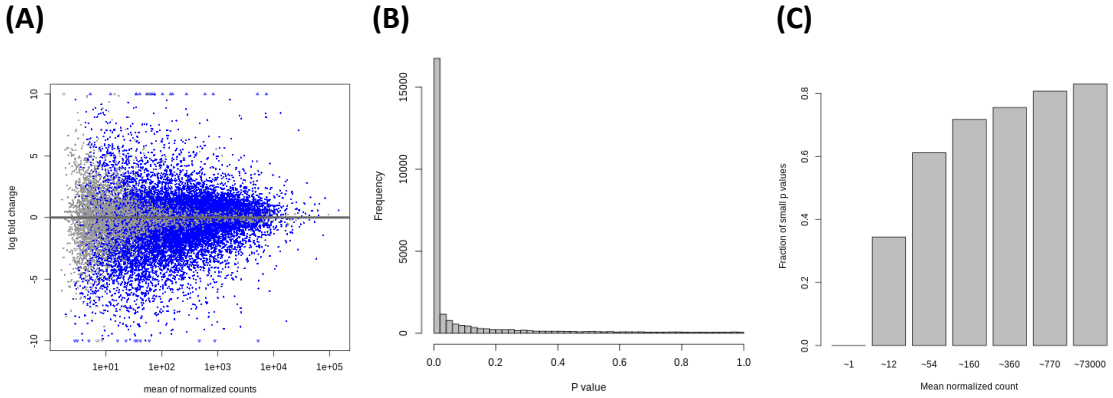
Supplementary Figure 3.4. Hidden batch effect identification by using full model matrix with the two factors days aftes anthesys (DAA) (A) and variety (B). Was estimated two surrogate variables for each factor considered in the constrast (variety and DAA). For the surrogate variables estimated for DAA and variety, the samples Sko_13_1 and Kan_13_3 respectively showed a slighly deviation.



Supplementary Figure 3.5. Representation of the standard deviation of count matrix gene expression against the mean before (A) and after (B) regularized-log transformation (rlog). As we can see here, the standard deviation is mean dependent before normalization in RNA-Seq counts experiments with a negative binomial distribution. This kind of transformatios was recommended to avoid the influence of hight expressed genes into the variance.

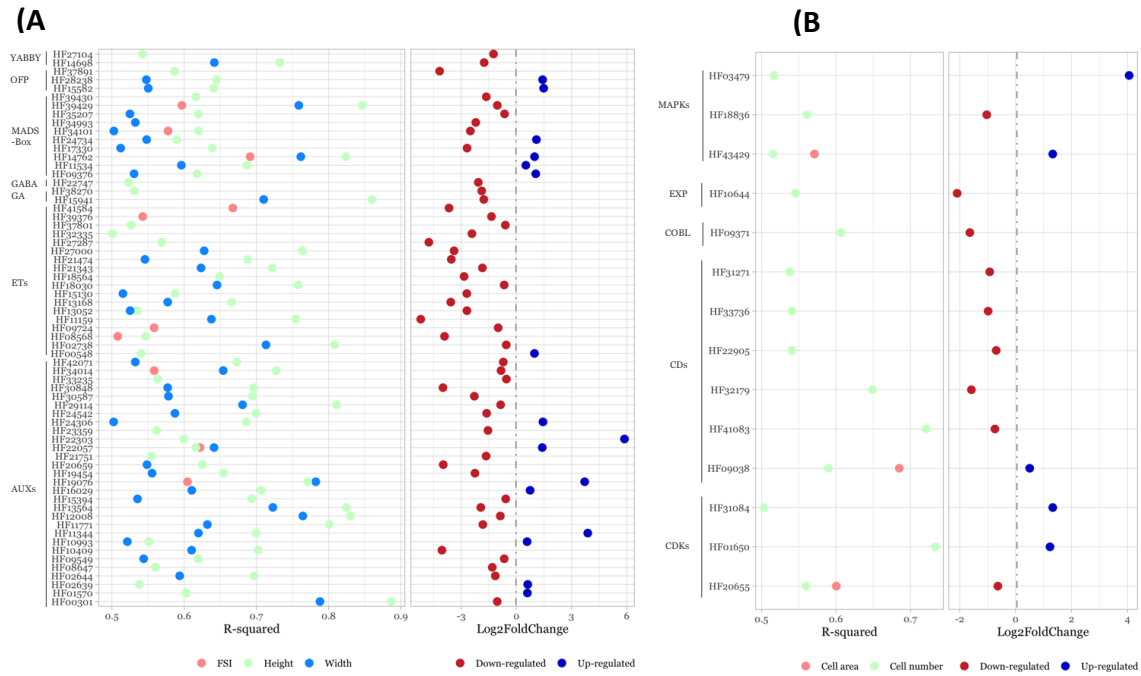


Supplementary Figure 3.6. Exploratory analysis and visualization of count data after deleted samples with high dissimilarity (Sko_13_1 and Kan_13_3). Count data matrix was pre-filtered to remove very low or no expressed genes and regularized-logarithm variance stabilizing transformation applied. A sample-to-sample Euclidean distance matrix heatmap (A) was plotted to assess overall similarity between samples, while a Principal Component Analysis (PCA) plot (B) project onto 2D plane the two principal components (PC) which capture most of the variance present in the dataset.



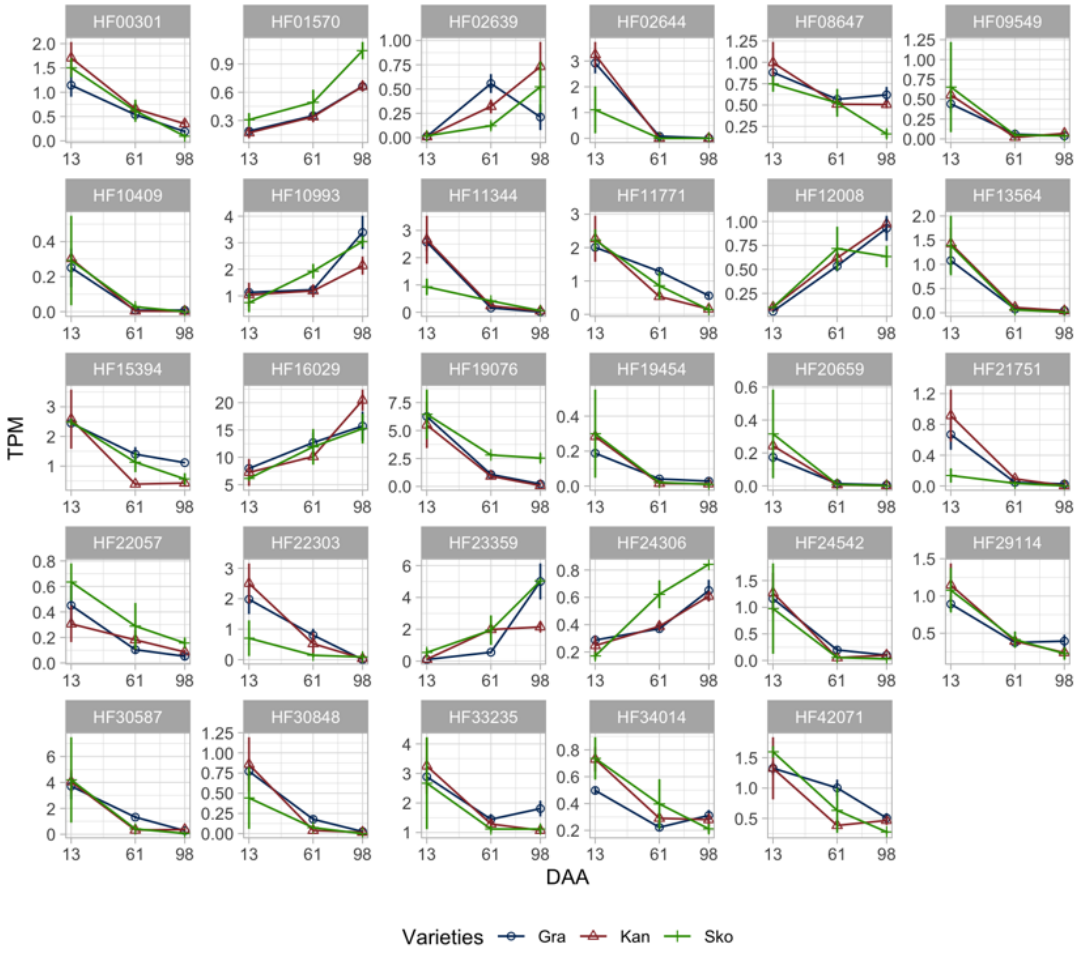
Supplementary Figure 3.7. Diagnostics plots obtained from time course analysis performed with the package DESeq2 over the apple varieties 'Gran'd Mere', 'Kansas Queen' and 'Skovfodge' at three diferent developmental stages 13, 61 and 98 DAA. (A) MA-plot provides a useful overview for the distribution of the estimated coefficient in the model by represent the log 2 fold change to mean of normalized counts. (B) *P* values histogram distribution plot gives us the view if the exclusion of very low or no expressed genes was done efficiently during pre-filtering the dataset before DEGs analysis. (C) Ratio of small *p* values for genes binned by mean normalized counts plot demonstrates

that genes with very low mean count have little or no power, and are best excluded from testing.

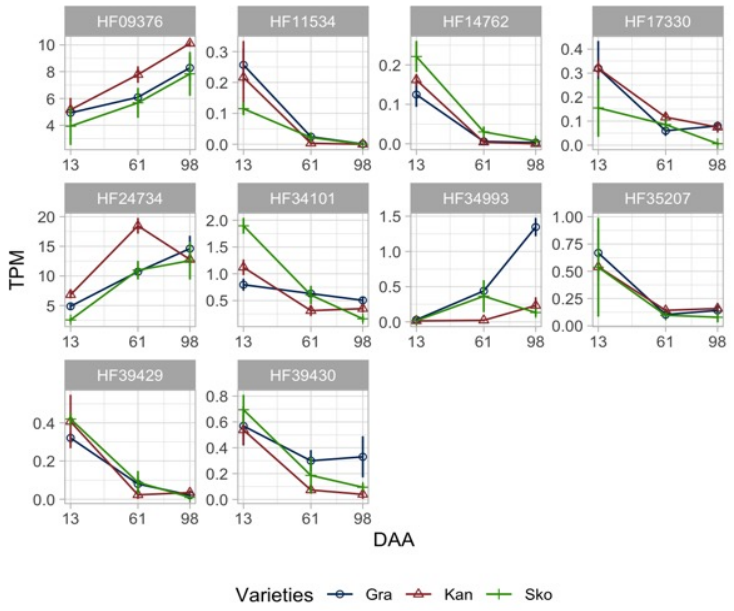


Supplementary Figure 3.8. Dot plot of R-squared and Log2FoldChange values of selected genes from TC-DEA results. **(A)** Values obtained from the analysis between the TPM count matrix and FSI, height and width (growth data) at points 13, 61 and 98 DAA. These genes are related to hormones (AUXs, ETs, GABA, GA) and regulators of leaf and fruit shape (YABBY, MADs-Box, OFP). **(B)** Values obtained from the analysis between the TPM count matrix and Cell area, cell number and intercellular spaces (microscopic analysis of cells from parenchyma tissue data) at points 61 and 98 DAA. These genes are related to cell division control (CDs, CDKs, COBL and MAPKs) and cell expansion (EXP).

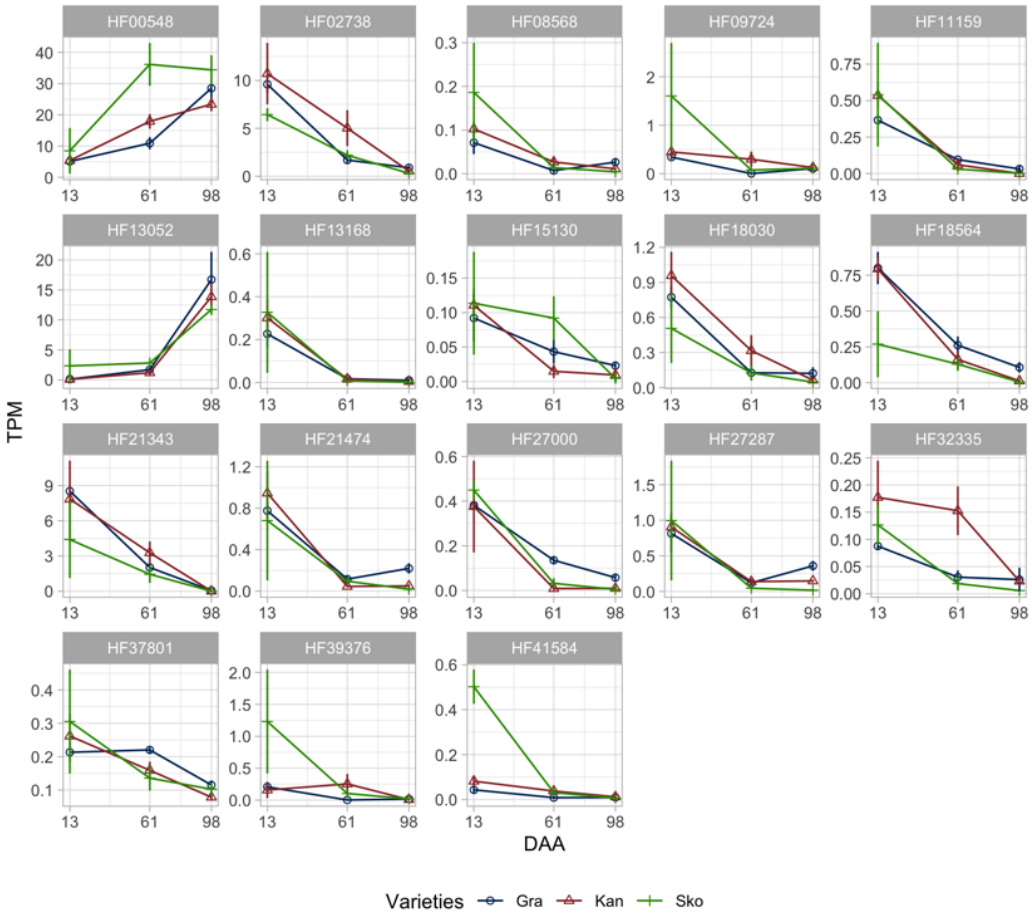
AUXs genes



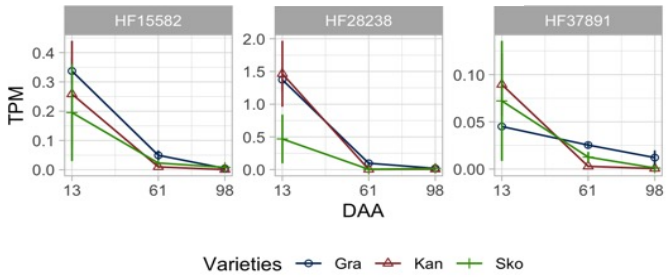
MADS-Box genes



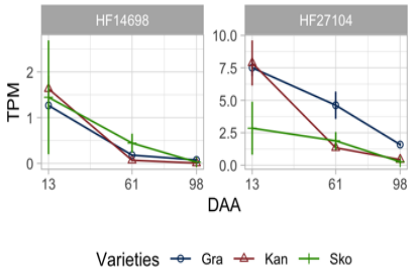
Ethylene genes



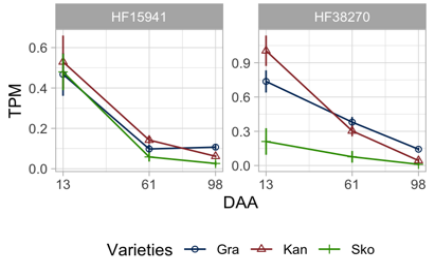
OFP genes



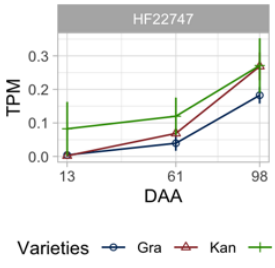
YABBY genes

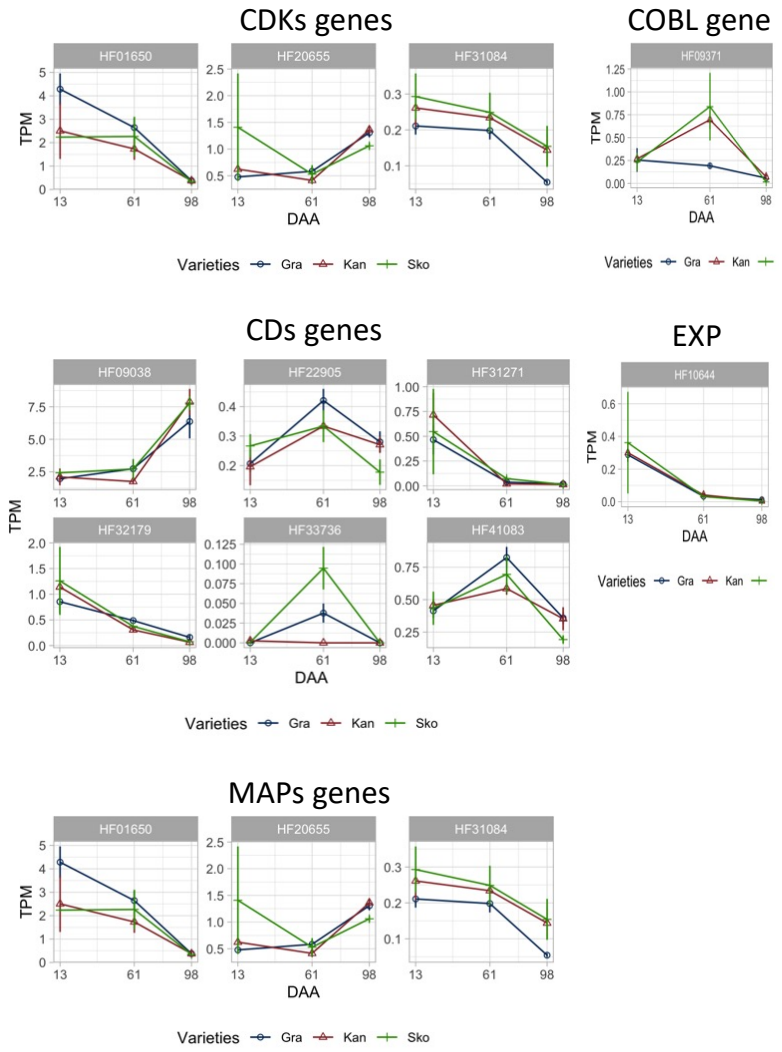


GA genes



GABA genes





Supplementray Figure 3.9. Lines plot of the comparative analysis of the count matrix normalized in TPM of the selected genes by gene annotations obtained in TC-DEA. Comparing the expression at stages 13, 61 and 98 DAA of the three varieties studied Grand'mere (GRA), Kansas Queen (KAN) and Skovfoged(SKO).

Supplementary Table 3.1. Record of dates from flowering to harvest of the genotypes selected for the growth.

	MUNQ	Name	Start flowering	Full flowering	Final flowering	Harvest	Total days
Oblate							
	2	Carrata	26/3/19	2/4/19	8/4/19	25/9/19	183
	653	Grand'mere	20/3/19	27/3/19	8/4/19	25/9/19	189
	33	Gros api	27/3/19	2/4/19	8/4/19	1/9/19	158
Spheroid							
	565	Kansas Queen	26/3/19	2/4/19	17/4/19	2/8/19	129
	1123	Horei	2/4/19	8/4/19	17/4/19	1/10/19	182
	2872	Pero dourado	27/3/19	2/4/19	8/4/19	28/8/19	154
Oblong							
	345	Skovfoged	20/3/19	2/4/19	18/4/19	8/10/19	202
	2784	Giambun	3/4/19	11/4/19	18/4/19	25/9/19	175
	12_O063	12_O063	27/3/19	2/4/19	18/4/19	1/10/19	188

Additional information: Trees planted: 15/04/2016

Supplementary Table 3.2. Count matrix summary obtained from reads featureCounts. Description of assigned and unassigned reads to annotated features (annotated exons in HFTH1 apple reference genome) before mapping reads and filtering. No reads were found unmapped, with mapping quality under MAPQ previously stablished (MAPQ < 30), assigned to chimaera genes, fragmented genes, duplicated regions, multiple mapping, overlapping, aligned as singletons, aligned to secondary sequences, splitted or ambiguously assigned.

Sample ID	Assigned Exons	Unassigned No features
Gra_13_1	17836315	2079187
Gra_13_2	19481308	2355795
Gra_13_3	20011554	2392307
Gra_61_1	20360425	2983308
Gra_61_2	18652163	2369291
Gra_61_3	18368452	2701191
Gra_98_1	21280688	2625478
Gra_98_2	19889778	2471230
Gra_98_3	24285257	2989522
Kan_13_1	16979451	2119040
Kan_13_2	21608007	2533643
Kan_13_3	20089306	2855454
Kan_61_1	15048910	2221388
Kan_61_2	18824544	2640411
Kan_61_3	21900987	2644375
Kan_98_1	20548094	2676339
Kan_98_2	21117364	2372440

Kan_98_3	19478883	2338390
Sko_13_1	19351603	2391442
Sko_13_2	21859530	2319074
Sko_13_3	26442009	2337302
Sko_61_1	19264201	1825725
Sko_61_2	21445845	2621650
Sko_61_3	19679688	3304304
Sko_98_1	20980113	2553848
Sko_98_2	18332678	2618135
Sko_98_3	18947903	2291263

Supplementary Table 3.3. Differentially expressed genes obtained from RNA-Seq samples from three apple fruit varieties ['Gran'd mere' as 'Gra'; 'Kansas Queen' as 'Kan', and 'Skovfodge' as 'Sko'], during fruit development: 13, 61 and 98 days after anthesys (DAA).

Contrast	DESeq2
Gra_13_vs_Kan_13	1845
Sko_13_vs_Kan_13	2451
Kan_61_vs_Kan_13	7067
Kan_98_vs_Kan_13	10487
Gra_61_vs_Gra_13	7658
Gra_98_vs_Gra_13	9474
Sko_61_vs_Sko_13	8968
Sko_98_vs_Sko_13	12587
Gra_13_vs_Sko_13	3049
Gra_61_vs_Kan_61	6491
Sko_61_vs_Kan_61	4590
Kan_98_vs_Kan_61	5991
Gra_98_vs_Gra_61	6109
Sko_98_vs_Sko_61	9018
Gra_61_vs_Sko_61	3918
Gra_98_vs_Kan_98	4083
Sko_98_vs_Kan_98	4806
Gra_98_vs_Sko_98	6746

MAIN DISCUSSION

Image-based morphometric analysis of apples

The appearance of the fruit subconsciously affects the consumer's perception of quality (Jaeger et al., 2018). It is the main sensory trait that consumers consider when evaluating fruit quality and make their purchase decisions (Ares et al., 2009) Fruit size and shape are among the most relevant traits in terms of attractiveness.

In this work we have exhaustively evaluated apple fruit size and shape through a wholistic approach combining phenotypes (fruit morphology, parenchyma organization) and genomic data (DNA and RNA) to ultimately identify genomic regions and suggest candidate genes for fruit size and shape regulation along development.

While the fruit size can be easily and unambiguously evaluated (through weight or metric dimensions), shape is a concept, a formal representation obtained through the intellective way. To homogenize criteria, examination, and evaluation offices (as the UPOV and the ECPGR) have released guides based on the experience of breeders and evaluators. Although these guides may suffice for the assignment of the fruits into few given classes, such descriptors are of less use for genomic studies.

With the aim of characterizing fruit shape considering multiple attributes, we started a high-throughput phenotypic analysis. The genotypes used were from the Apple REFPOP collection, formed by cultivars and seedlings representative of the apple variability in Europe (Jung et al., 2020; 2022). In total we cut in halves close to 6,500 apples, which were scanned. This dataset of images constitutes a highly valuable tool for further analysis. Currently the use of images in plant phenotyping, combined artificial

intelligence methods and mathematical models, surges as a promising strategy to assist breeders and scientists in the study and prediction of desired traits. For example, Recently Chakrabarti et al., (2021) applied mathematical models to mimic apple growth. In peach Cirilli et al., (2021) used fruit images to identify QTLs for fruit size and shape.

So far, the images obtained here have been used in the frame of the European project INVITE and in collaboration with Dr. David Rousseau team (University of Angers) to develop a method based on computer vision and unsupervised machine learning to automatically classify apples into given shapes (Mouad et al. under revision in Biosystems Engineering) and to develop a web application, called PanoVar, to manually classify apple images into shape classes (Mouad et al. to be released). Also, the images are being used by other researchers of the team to develop new tools and novel genome prediction models.

Here we processed the images with the Tomato Analyzer V3 software and obtained data on 15 attributes of size and shape. The Tomato Analyzer software was developed to evaluate tomatoes but can be applied to take fruit measurements in several species (Gonzalo et al., 2009; Nankar et al., 2020; Pereira et al., 2021; Sierra-Orozco et al., 2021). Its use to evaluate apple sections was laborious since it does not recognize correctly the apple contour, requiring visual inspection of all images and manual adjustments of most of them. Despite this inconvenience, we obtained accurate measurements of size and shape attributes. For shape, we recorded measurements describing the angle of opening in the peduncular cavity and eye basin (PAMa and DAMa), shoulder height measurement (DFB and PFB) and its relation (FST). As well as height/diameter ratio (FSII) and calculating FSII based on the measurement of eccentricity (FSIINT).

Despite the relevance of fruit size and shape, only few works study their heritability and variation along fruit evolution. Here we used data obtained in three harvest seasons to find out that, in general, size traits had higher heritability than shape traits. The FSII ration was the attribute with higher heritability (0.82). This is consistent with the high heritability found for this trait (0.79) by Currie et al., (2000).

The study of apple shape along fruit development (**Chapter 3, Figure 3.1**) shows that FSII is already a shape discriminant parameter at the early stages of development and keeps evolving towards the final shape along cell division and cell expansion stages.

While the FSI is the parameter par excellence in shape description, the relevance of other traits in determining fruit shape is poorly known. We have used the machine learning tool to test which parameters are the most important for apple shape. In the last decade, machine learning tools have already been implemented to identify and predict events in agriculture (Meshram et al., 2021), such as for characterization and selection of interesting genetic resources in a breeding program (Danckaers et al., 2017). One of these tools is the random forest algorithm used in this work As Random Forest relays in the construction of multiple decision trees, it retains their advantages while using grouped samples, random variable subsets to achieve better results, and handles missing values. As well as allow to use of several types of variables (continuous, binary, and categorical), it is suitable for modeling multidimensional data (Qi, 2012). RF has been used in crops to take decisions in several biological applications (Sánchez-Galán et al., 2021; Moradi et al., 2021; Zhang et al., 2021). In our study, this algorithm with classification supervised by hundreds of estimators made accurate predictions of visual categories with specific measures, in both categories is 0.90, but for the F1 score in flat

and spherical shape, the prediction was moderately low. These two classes were difficult to differentiate visually. While we have evaluated the error of the model, we did not evaluate human error. Finally, the supervised machine learning identified found that the FSII and FST measures were the more relevant for the fruit assignment into classes, followed by the distal angle Macro (DAMa), the eccentricity (ECC), and the proximal angle macro (PAMa). These parameters should be added in future apple shape analyzer software for fruit evaluation and classification.

Mapping for shape and size measures in apple

Using the measurements described in **Chapter 1**, we performed a genome-wide association study (GWAS) to search for genomic regions controlling size and shape attributes. Several studies for fruit shape and size have been reported (Kenis et al., 2008; Chang et al., 2014; Potts et al., 2014; Cao et al., 2015; Jung et al., 2022). In addition, few genes have been suggested to be involved in fruit size and shape determination (Yao et al., 2015 and Yao et al., 2018), the knowledge of underlaying genomic loci remains limited.

We found association for all but FST, DAMa, E and ECC data. The heritability of these traits had values below 40%) (Supplementary table 1.4) which may explain the absence of associations.

In this study, we have identified 71 QTNs for 11 shape and size attributes. With these identified markers and those associated with size and shape traits already published (Kenis et al., 2008; Chang et al., 2014; Potts et al., 2014; Cao et al., 2015; Jung et al., 2022)., we have constructed a PhenoGram detailing the position of a physical map of all

these markers (110). One of the chromosomes with the higher number of markers (19) described was chromosome 11. In this chromosome we found two of 11 SNPs, associated with FSIINT and CAT-own measurements (related to fruit shape) in a 9.7 kb haploblock on top of the chromosome. The cultivars 'Grand'mere' (flat and large) and 'Skovfoged' (oblong shape) were homozygous for both SNPs, while 'Kansas Queen' (round shape) was heterozygous. Cao et al., (2015) reported a QTL for FSII, at 5 Mb of these two SNPs (AX-115327898 and AX-115327900). Chang et al., (2014) also detected several QTLs for fruit shape index (FSI), one of those QTLs in LG11 contributed to a phenotypic variance between 10.3 - 13.7% in a segregating population.

Two ovate family protein genes (*MdOFP17* and *MdOFP4*) are annotated in this region where two SNP markers have been found highly associated to FSIINT and CAT-down traits. These two traits are highly correlated, as was observed in the correlation analysis and explained by the Random forest analysis (both in **Chapter 1**). As we have seen, the fruit shape index (FSII) measure is the most considered in shape studies. Our results in Chapter 1 confirm its relevance as it has the greatest weight in the definition of fruit shape.

In other species as in tomato, several genes for the control of fruit shape have been identified, such as *OVATE*, *SUN*, *FAS* (fasciated) and *LC* (Locule number). The *SUN* and *OVATE* genes control shape elongation, while *FAS* and *LC* control locule number and flat shape (Tanksley, 2004; Brewer et al., 2007; Rodriguez et al., 2011). In pepper the fw2.1 locus co-localizes with the ovate gene and is associated with smaller fruit (Zygier et al., 2005), also in cucumber, three QTLs for fruit shape have been identified, one of them (fruit shape index 2.1) with a phenotypic variation greater than 50% (Gao et al., 2020).

In melon, the QTL *fsqs8.1* is associated with the round shape of the fruit, at whose locus the gene *CmOFP13* (ovate family protein) is annotated (Martinez-Martinez et al., 2022). In peach, a 1.7 Mb downstream inversion of the gene encoding the ovate *PpOFP1* is responsible for the flat shape (Zhou et al., 2021).

Also, QTNs for size have been identified on chromosome 11. One of them, the QTN for fruit height was at 216 kb distance from the SNP AX-115464400 in Jung et al., (2022). Yao et al., (2015), identified a microRNA (miRNA172) in a QTL for fruit size inn chromosome 11 which overexpression influences fruit size.

Another genomic region that could be interesting because of its annotated genes is the QTN for the Circular measure in an 18.7 kb haploblock on chromosome 13. The TCP15-like transcription factor is the only gene in the block. This gene is involved in the regulation of plant development and in the stimulation of biosynthesis of hormones such as brassinosteroid, jasmonic acid, and flavonoids (Li, 2015).

Two QTNs (for A and WHM) were identified on chromosome 5, close to two QTLs already described (Chang et al., 2014 and kenis et al., 2008). Both SNPs jointly explain 6.57 % of the phenotypic variance. According to the TAIR database description, the annotated genes in these regions are responsible for growth regulation, such as Transcriptional factor B3 family protein/auxin-responsive factor AUX/IAA-related (HF12008), ethylene responsive element binding factor 1 (HF11991), Gibberellin-regulated family protein (HF08230) and Auxin-responsive GH3 family protein (HF08237). Hormones play an important role in fruit growth (Kumar et al., 2013) and are controlled by multiple genes. In melon, two overlapping QTLs, one for fruit diameter and one for fruit weight were

detected on chromosome 11, identifying the gene *MELO3C025758* (auxin response factor) as one of the candidate genes for these traits (Lian et al., 2021). In tomato, auxin and gibberellin hormones regulate the transition from flower stage to fruit set (Jong et al., 2009). Endogenous auxin concentration is one of the factors controlling fruit size in apple (Bu et al., 2020). Devoghalaere et al., (2012) suggest a potential role in apple fruit size of the Auxin Responsible Factor (ARF106) gene, identified in a QTL for fruit weight on chromosome 15.

A possible role of these candidate genes in determining the natural variation of apple size and shape will need to be validated. They could be validated through CRISPR/Cas9-based genome editing, which is feasible in apple (Malnoy et al., 2016), or by ectopic expression as done in Yao et al., (2018).

Fruit development and differentially expressed genes

To add more information to the morphometric characterization, we decided to determine the dynamics of fruit growth and shape formation along development in terms of the morphology parameters as well as in terms of the histological structure of the parenchyma. To select the developmental stages and cultivars for such analysis, we observed the variation of the measures acquired in Chapter 1 along fruit development in 9 cultivars of three fruit typologies. From this data we selected three cultivars (the flat 'Grand' mere' GRA, the round 'Kansas queen' KAN, and the oblong 'Skovfoged' SKO) and three developmental stages (0, 61 and 98 DAA).

At 0 DAA we found differences in the width and height of the ovary in the three genotypes. The GRA genotype shows greater width and a flattened appearance in height, in KAN both measurements are similar, and in contrast to SKO it shows an elongation in the apical zone. From 0 to 61 DAA, FSI reduces for all fruit shape typologies.

At 61 DAA FSI variation becomes slower, till finally reach a plateau at 98 DAA (**Chapter 3, Figure 1**).

According to the correlation values between the morphometric and histological measures, fruit size (height and width) and cell area have a direct relation.

To put some light in the molecular mechanisms underlying fruit shape and size, we incorporated genomic data to the study: RNAseq, whole genome DNA sequences and whole genome bisulfite DNA sequences. While whole genome DNA was extracted from leaves, total RNAseq was extracted from the hypanthium of fruits at three development stages: 13, 61 and 98 DAA.

We have performed numerous differential gene expression contrasts between the varieties, at the three developmental stages, and along development (time-course differential expression analysis (TC-DEA)).

We calculated the correlation between the transcript levels of the genes identified and, in the TC-DEA annotated genes and the traits (morphometric and histologic) and calculated the coefficient of determination, the R^2, which expresses the proportion variance of the traits explained (PVE) by the transcript levels (transcript per million, TPMs). Among the genes identified, multiple phytohormones explained more than 50 % of the variance (PVE). This is the case of auxins, known to be involved in the regulation of various aspects of fruit development such as cell proliferation, cell expansion and fruit ripening (Srivastava and Handa, 2005). In one study, auxins were determined as responsible for the final size of the apple (Devoghalaere et al., 2012). In this analysis, the ARF9 gene (HF08647, PREDICTED: auxin response factor 9) was identified and associated with fruit height along growth and with down-regulated activity. According to Wang et al., (2005) and de Jong et al., (2015) tomatoes overexpressing the ARF9 gene reduced

fruit size and had a down-regulated activity on cell production during early fruit development. Homologous to IAA (indole-3-acetic acid), which are known to be involved in the regulation of auxin-mediated gene expression, were also identified along fruit growth. In tomato, the downregulation of such genes results in fruit development without need of pollination and fertilization Wang et al., (2005). In Devoghalaere et al., (2012), IAA genes increase the cortex or hypanthium zone of the fruit.

Another phytohormone identified in the TC-DEA was ethylene (ET), known as responsible for the ripening of several fruit species such as melon, tomato, apple (Pereira et al., 2020; Liu et al., 2016; Yue et al., 2020). We identified fifteen ethylene-related proteins, one of them is the gene (HF13168, Predicted: AP2-like ethylene-responsive transcription factor ANT) with down-regulated activity involved in the control of primary and secondary metabolism in growth and development, as well as in responses to environmental stimulation (Licausi et al., 2013).

The phytohormone gibberellin (GA) was also identified in genes differentially expressed during growth with a PVE>0.53 in relation to fruit height and width, such as the gene (HF15941: *gibberellin receptor Gid1C-like*), described as nuclear GA insensitive dwarf1s (GID1s) receptors responsible for triggering degradation of DELLAs repressors. In Arabidopsis at the early stage of fruit development they are transcriptionally active and play an important role in seed development and pod elongation (Gallego-Giraldo et al. 2014). In apple, GA applications at the fruit set stage induce fruit shape change, showing a greater growth in both height and width (Nakagawa et al. 1968).

In addition to the hormones already mentioned, other hormones with lower PVE values were also identified, such as jasmonates related to the fruit ripening process (Li et al. 2017), brassinosteroids promoting cell proliferation, fruit ripening and senescence

(Clouse, 2011), and abscisic acid (ABA) involved in fruit set abscission, but also found an endogenous concentration of ABA in the fruit cortex (Eccher et al., 2013).

Genes of the *MADs-Box* protein family are also linked to development and growth. We found nine in the TC-DEA. In apple, genes of the *APETALA1* (*AP1*) and *AGAMOUS* (*AG*) group showed differential expression in the core, cortex and skin in young fruits (Yao et al., 1999).

When considering histological parameters, filtered values of PVE>0.5 identified genes controlling cell division and expansion, such as the *MAPKKK* (*mitogen-activated protein kinase kinase kinase*) gene cluster with function in the transduction of environmental and developmental signals, in addition to cell cycle progression (Jagodzik et al., 2018). One of them is *NPK1*, which is associated in a higher percentage of PVE with cell area and up-regulated activity. According to Nishihama et al., (2001) it has activity in the M phase of cell division, which is essential for the formation of the cell plate and its lateral growth, and therefore is required for cytokinesis. In relation to cell expansion, nine EXPANSINs (EXP) genes have been identified. These proteins are known to have a loosening activity, cell expansion and cell wall modification (Sampredro and Cosgrove, 2005). One of them is *MdEXPA20* associated with cell number down-regulated activity in the TC-DEA. In a study in apple, Zhang et al., (2014) found that *MdEXPA20* expression plays an important role in fruit development in relation to cell expansion during growth.

In summary, the contrats and time-course studies for gene differential expression analysis identified hundreds of genes. Among them we focused on those involved in the regulation of hormones because of their already known role in organ development. To search for candidate genes outside the phytohormones we explored the differential expression patterns of the candidate genes found in the GWAS analysis, and therefore

putatively responsible for the natural variation of apple shape and size. Some of the genes picked from the GWAS results were also hormone-related and had been already selected from the general analysis. Others (mainly transcription factors) were found to be differentially in one or more contrasts.

We found specially interesting the gene HF43536, annotated as *OFP4-like* transcription repressor, which was found associated to FSII. This gene is not expressed in the cultivar with oblong fruits (SKO), while is highly expressed in the flat GRA and expressed, although with lower levels of transcripts, in the round KAN. DNA sequence analysis revealed a structural variation in the promoter of the gene. Further analyses are required to validate the polymorphisms (only determined *in silico* in this work) and its association in cultivars and progenies.

In whole, this exhaustive and holistic work, combining morphometric, histologic, genetic, and genomic analysis, contributes enormously to increase the knowledge on the genetics and genomics behind apple fruit shape and size determination. Moreover, we have developed relevant tools and data that will help for further studies. We have generated a large set of images that will be of great use to develop a shape analyzer software for apple as well as for future genomic studies. So far, these images have been already used for automatic and manual cultivar classification. In addition, we provide a list of genes related to fruit shape and size variation to be validated. We also provide a large data set of RNA sequences obtained along fruit development, and last but not least, a marker for fruit shape that, in case of validated, may help to select or discard, or characterize, oblate apples from the DNA analysis.

CONCLUSIONS

1- The Tomato Analyzer software provided good morphometric data of 15 size and shape attributes acquired from 2D apple images. However, for the analysis the images required visual inspection and manual correction, limiting its use for high-throughput phenotyping in apple.

2- Statistical analysis showed strong correlation between the size attributes, while shape attributes were low to moderately correlated. In general, size traits had higher heritability than shape traits (0.72 *vs* 0.45 in average, respectively). Among all the parameters evaluated, FSII was the one more determinant in fruit classification followed by FST, DAMa, ECC and PAMa. This information will be relevant for breeding, cultivar identification, and will help in the design of a software for apple shape analysis.

3- The GWAS analysis using two methods (FarmCPU and BLINK) in four datasets, corresponding to three years of data and to the mean of the values, identified 59 SNPs associated with fruit size and shape traits (35 with FarmCPU and 45 with BLINK) responsible for 71 QTNs. The QTNs were distributed in all chromosomes but in chromosome 10 and 15.

4- Thirty-four QTNs, identified by 27 SNPs, were related for size traits. Some of the QTNs for fruit area, width at mid height co-localized or mapped at close distance. Strong QTNs were found in chromosome 2.

5- Thirty-seven QTNs, identified by 26 SNPs, were related to shape attributes. Nine were distributed along chromosome 11. A haploblock of 9.7 kb in this chromosome associated to FSIINT and CAT-own data contains two genes of the

ovate protein family (*MdOFP17* and *MdOFP4*), described in other works for their role in fruit shape determination.

6- In the genomic regions of the QTNs identified for size measurements (MH and MW), we found hormone related genes that are reported to play an important role in fruit growth.

7- Along fruit development, we found differences in cell number, area and intercellular spaces when compared the hypanthium of three cultivars with different shape (flat, round and ovate)

8- Differential gene expression analysis between the varieties at three developmental stages (13, 61 and 98 DAA) and along development identified multiple genes with PVE over 50%, such as *ARF9*, *SAUR15*, *Gid1C*-like, *NPK1*, *MdEXPA20* .

9- The *MdOFP4* gene stood out in all analysis, rising as candidate for fruit shape in apple. A polymorphism in the promoter of this gene may be involved in its lack of expression in the oblong genotype.

10- This work contributes enormously to increase the knowledge on the genetics and genomics behind apple fruit shape and size determination. Moreover, we have developed relevant tools and data that will help for further studies on fruit shape and development. We have generated a large set of images that will be of great use to develop a shape analyzer software for apple as well as for future genomic studies.

MAIN BIBLIOGRAPHY

1. Albacete, A., Martínez-Andújar, C., Martínez-Pérez, A., Thompson, A. J., Dodd, I. C., & Pérez-Alfocea, F. (2015). Unravelling rootstock× scion interactions to improve food security. *Journal of experimental botany*, *66*(8), 2211-2226.

2. Amyotte, B., Bowen, A. J., Banks, T., Rajcan, I., & Somers, D. J. (2017). Mapping the sensory perception of apple using descriptive sensory evaluation in a genome wide association study. *PloS one*, *12*(2), e0171710.

3. Ares, G., Barrios, S., Lareo, C., & Lema, P. (2009). Development of a sensory quality index for strawberries based on correlation between sensory data and consumer perception. *Postharvest biology and technology*, *52*(1), 97-102.

4. Ben Sadok, I., Tiecher, A., Galvez-Lopez, D., Lahaye, M., Lasserre-Zuber, P., Bruneau, M., ... & Laurens, F. (2015). Apple fruit texture QTLs: year and cold storage effects on sensory and instrumental traits. *Tree genetics & genomes*, *11*(6), 1-20.

5. Bianco, L., Cestaro, A., Linsmith, G., Muranty, H., Denancé, C., Théron, A., ... & Troggio, M. (2016). Development and validation of the Axiom® Apple480K SNP genotyping array. *The Plant Journal*, *86*(1), 62-74.

6. Bianco, L., Cestaro, A., Sargent, D. J., Banchi, E., Derdak, S., Di Guardo, M., ... & Troggio, M. (2014). Development and validation of a 20K single nucleotide polymorphism (SNP) whole genome genotyping array for apple (Malus× domestica Borkh). *PloS one*, *9*(10), e110377.

7. Brewer, M. T., Moyseenko, J. B., Monforte, A. J., & van der Knaap, E. (2007). Morphological variation in tomato: a comprehensive study of quantitative trait loci controlling fruit shape and development. *Journal of experimental botany*, *58*(6), 1339-1349.

8. Bu, H., Yu, W., Yuan, H., Yue, P., Wei, Y., & Wang, A. (2020). Endogenous auxin content contributes to larger size of apple fruit. *Frontiers in plant science*, *11*, 592540.

9. Bump, V. L. (1989). Apple pressing and juice extraction. In *Processed apple products* (pp. 53-82). Springer, New York, NY.

10. Bus, V. G., Laurens, F. N., Van De Weg, W. E., Rusholme, R. L., Rikkerink, E. H., Gardiner, S. E., ... & Plummer, K. M. (2005). The Vh8 locus of a new gene-for-

gene interaction between Venturia inaequalis and the wild apple Malus sieversii is closely linked to the Vh2 locus in Malus pumila R12740-7A. *New Phytologist*, *166*(3), 1035-1049.

11. Cao, K., Chang, Y., Sun, R., Shen, F., Wu, T., Wang, Y., ... & Han, Z. (2015). Candidate gene prediction via quantitative trait locus analysis of fruit shape index traits in apple. *Euphytica*, *206*(2), 381-391.

12. Cappellin, L., Costa, F., Aprea, E., Betta, E., Gasperi, F., & Biasioli, F. (2015). Double clustering of PTR-ToF-MS data enables the mapping of QTLs related to apple fruit volatilome. *Scientia Horticulturae*, *197*, 24-32.

13. Cappellin, L., Farneti, B., Di Guardo, M., Busatto, N., Khomenko, I., Romano, A., ... & Costa, F. (2015). QTL analysis coupled with PTR-ToF-MS and candidate gene-based association mapping validate the role of Md-AAT1 as a major gene in the control of flavor in apple fruit. *Plant Molecular Biology Reporter*, *33*(2), 239-252.

14. Chagné, D., Crowhurst, R. N., Troggio, M., Davey, M. W., Gilmore, B., Lawley, C., ... & Peace, C. (2012). Genome-wide SNP detection, validation, and development of an 8K SNP array for apple. *PloS one*, *7*(2), e31745.

15. Chagné, D., Kirk, C., How, N., Whitworth, C., Fontic, C., Reig, G., ... & Iglesias, I. (2016). A functional genetic marker for apple red skin coloration across different environments. *Tree Genetics & Genomes*, *12*(4), 1-9.

16. Chagné, D., Vanderzande, S., Kirk, C., Profitt, N., Weskett, R., Gardiner, S. E., ... & Bassil, N. V. (2019). Validation of SNP markers for fruit quality and disease resistance loci in apple (Malus× domestica Borkh.) using the OpenArray® platform. *Horticulture research*, *6*.

17. Chakrabarti, A., Michaels, T. C., Yin, S., Sun, E., & Mahadevan, L. (2021). The cusp of an apple. *Nature Physics*, *17*(10), 1125-1129.

18. Chang, Y., Sun, R., Sun, H., Zhao, Y., Han, Y., Chen, D., ... & Han, Z. (2014). Mapping of quantitative trait loci corroborates independent genetic control of apple size and shape. *Scientia Horticulturae*, *174*, 126-132.

19. Chen, X., Li, S., Zhang, D., Han, M., Jin, X., Zhao, C., ... & An, N. (2019). Sequencing of a wild apple (Malus baccata) genome unravels the differences

between cultivated and wild apple species regarding disease resistance and cold tolerance. *G3: Genes, Genomes, Genetics*, *9*(7), 2051-2060.

20. Christeller, J. T., McGhie, T. K., Johnston, J. W., Carr, B., & Chagné, D. (2019). Quantitative trait loci influencing pentacyclic triterpene composition in apple fruit peel. *Scientific Reports*, *9*(1), 1-7.

21. Cirilli, M., Baccichet, I., Chiozzotto, R., Silvestri, C., Rossini, L., & Bassi, D. (2021). Genetic and phenotypic analyses reveal major quantitative loci associated to fruit size and shape traits in a non-flat peach collection (P. persica L. Batsch). *Horticulture Research*, *8*.

22. Clouse, S. D. (2011). Brassinosteroid signal transduction: from receptor kinase activation to transcriptional networks regulating plant development. *The Plant Cell*, *23*(4), 1219-1230.

23. Cornille, A., Giraud, T., Bellard, C., Tellier, A., Le Cam, B., Smulders, M. J. M., ... & Gladieux, P. (2013). Postglacial recolonization history of the E uropean crabapple (Malus sylvestris M ill.), a wild contributor to the domesticated apple. *Molecular Ecology*, *22*(8), 2249-2263.

24. Cornille, A., Giraud, T., Smulders, M. J., Roldán-Ruiz, I., & Gladieux, P. (2014). The domestication and evolutionary ecology of apples. *Trends in Genetics*, *30*(2), 57-65.

25. Cornille, A., Gladieux, P., Smulders, M. J., Roldán-Ruiz, I., Laurens, F., Le Cam, B., ... & Giraud, T. (2012). New insight into the history of domesticated apple: secondary contribution of the European wild apple to the genome of cultivated varieties. *PLoS genetics*, *8*(5), e1002703.

26. Costa, F., Cappellin, L., Zini, E., Patocchi, A., Kellerhals, M., Komjanc, M., ... & Biasioli, F. (2013). QTL validation and stability for volatile organic compounds (VOCs) in apple. *Plant Science*, *211*, 1-7.

27. Costa, F., Stella, S., de Weg, V., Eric, W., Guerra, W., Cecchinel, M., ... & Sansavini, S. (2005). Role of the genes Md-ACO1 and Md-ACS1 in ethylene production and shelf life of apple (Malus domestica Borkh). *Euphytica*, *141*(1), 181-190.

28. Costes, E., & García-Villanueva, E. (2007). Clarifying the effects of dwarfing rootstock on vegetative and reproductive growth during tree development: a study on apple trees. *Annals of Botany*, *100*(2), 347-357.

29. Currie, A. J., Ganeshanandam, S., Noiton, D. A., Garrick, D., Shelbourne, C. J. A., & Oraguzie, N. (2000). Quantitative evaluation of apple (Malus× domestica Borkh.) fruit shape by principal component analysis of Fourier descriptors. *Euphytica*, *111*(3), 221-227.

30. Daccord, N., Celton, J. M., Linsmith, G., Becker, C., Choisne, N., Schijlen, E., ... & Bucher, E. (2017). High-quality de novo assembly of the apple genome and methylome dynamics of early fruit development. *Nature genetics*, *49*(7), 1099-1106.

31. Danckaers, F., Huysmans, T., Dael, M. V., Verboven, P., Nicolaï, B., & Sijbers, J. (2017). Building 3D statistical shape models of horticultural products. *Food and bioprocess technology*, *10*(11), 2100-2112.

32. De Jong, M., Wolters-Arts, M., Schimmel, B. C., Stultiens, C. L., de Groot, P. F., Powers, S. J., ... & Rieu, I. (2015). Solanum lycopersicum AUXIN RESPONSE FACTOR 9 regulates cell division activity during early tomato fruit development. *Journal of experimental botany*, *66*(11), 3405-3416.

33. De Witte, K.; Vercammen, J.; van Daele, G.; Keulemans, J. (1996). Fruit set, seed set and fruit weight in apple as influenced by emasculation, self-pollination and cross-pollination. Acta Horticulturae, (423), 177–184.

34. Dennis, F. J. (2003). Flowering, pollination and fruit set and development. In *Apples: botany, production and uses* (pp. 153-166). Wallingford UK: CABI Publishing.

35. Devoghalaere, F., Doucen, T., Guitton, B., Keeling, J., Payne, W., Ling, T. J., ... & David, K. M. (2012). A genomics approach to understanding the role of auxin in apple (Malus x domestica) fruit size control. *BMC Plant Biology*, *12*(1), 1-15.

36. Dickinson, J. P., & White, A. G. (1986). Red colour distribution in the skin of Gala apple and some of its sports. *New Zealand journal of agricultural research*, *29*(4), 695-698.

37. Domínguez, E., Cuartero, J., & Heredia, A. (2011). An overview on plant cuticle biomechanics. *Plant Science*, *181*(2), 77-84.

38. Duan, N., Bai, Y., Sun, H., Wang, N., Ma, Y., Li, M., ... & Chen, X. (2017). Genome re-sequencing reveals the history of apple and supports a two-stage model for fruit enlargement. *Nature Communications*, *8*(1), 1-11.

39. Eccher, G., Alessandro, B., Mariano, D., Andrea, B., Benedetto, R., & Angelo, R. (2013). Early induction of apple fruitlet abscission is characterized by an increase of both isoprene emission and abscisic acid content. *Plant physiology*, *161*(4), 1952-1969.

40. Eccher, G., Ferrero, S., Populin, F., Colombo, L., & Botton, A. (2014). Apple (Malus domestica L. Borkh) as an emerging model for fruit development. *Plant Biosystems-An International Journal Dealing with all Aspects of Plant Biology*, *148*(1), 157-168.

41. Ellegren, H., & Galtier, N. (2016). Determinants of genetic diversity. *Nature Reviews Genetics*, *17*(7), 422-433.

42. Elsysy, M. A., & Hirst, P. M. (2017). The role of spur leaves, bourse leaves, and fruit on local flower formation in apple: An approach to understanding biennial bearing. *HortScience*, *52*(9), 1229-1232.

43. Evans, R. C., & Campbell, C. S. (2002). The origin of the apple subfamily (Maloideae; Rosaceae) is clarified by DNA sequence data from duplicated GBSSI genes. *American journal of botany*, *89*(9), 1478-1484.

44. FAOSTAT. (2020). Food and agriculture Organization of the United Nations. Available at: https://www.fao.org/faostat [Accessed October 15, 2022].

45. Fazio, G. (2021). Genetics, breeding, and genomics of apple rootstocks. In *The apple genome* (pp. 105-130). Springer, Cham.

46. Fazio, G., Aldwinckle, H. S., Volk, G. M., Richards, C. M., Janisiewicz, W. J., & Forsline, P. L. (2007, September). Progress in evaluating Malus sieversii for disease resistance and horticultural traits. In *XII EUCARPIA Symposium on Fruit Breeding and Genetics 814* (pp. 59-66).

47. Fazio, G., Chao, C. T., Forsline, P. L., Richards, C., & Volk, G. (2012, December). Tree and root architecture of Malus sieversii seedlings for rootstock

breeding. In *X International Symposium on Integrating Canopy, Rootstock and Environmental Physiology in Orchard Systems 1058* (pp. 585-594).

48. Fischer, C. (1994). Shortening of the juvenile period in apple breeding. In *Progress in temperate fruit breeding* (pp. 161-164). Springer, Dordrecht.

49. Forshey, C. G., Elfving, D. C., & Stebbins, R. L. (1992). *Training and pruning apple and pear trees*. American Society for Horticultural Science.

50. Gallego-Giraldo, C., Hu, J., Urbez, C., Gomez, M. D., Sun, T. P., & Perez-Amador, M. A. (2014). Role of the gibberellin receptors GID 1 during fruit-set in Arabidopsis. *The Plant Journal*, *79*(6), 1020-1032.

51. Gao, Y., Liu, F., Wang, K., Wang, D., Gong, X., Liu, L., ... & Volk, G. M. (2015). Genetic diversity of Malus cultivars and wild relatives in the Chinese National Repository of Apple Germplasm Resources. *Tree genetics & genomes*, *11*(5), 1-9.

52. Gao, Z., Zhang, H., Cao, C., Han, J., Li, H., & Ren, Z. (2020). QTL mapping for cucumber fruit size and shape with populations from long and round fruited inbred lines. *Horticultural Plant Journal*, *6*(3), 132-144.

53. Gapper, N. E., Rudell, D. R., Giovannoni, J. J., & Watkins, C. B. (2013). Biomarker development for external CO2 injury prediction in apples through exploration of both transcriptome and DNA methylation changes. *AoB Plants*, *5*.

54. Gianfranceschi, L., & Soglio, V. (2003, September). The European project HiDRAS: innovative multidisciplinary approaches to breeding high quality disease resistant apples. In *XI Eucarpia Symposium on Fruit Breeding and Genetics 663* (pp. 327-330).

55. Gonzalo, M. J., Brewer, M. T., Anderson, C., Sullivan, D., Gray, S., & van der Knaap, E. (2009). Tomato fruit shape analysis using morphometric and morphology attributes implemented in Tomato Analyzer software program. *Journal of the American Society for Horticultural Science*, *134*(1), 77-87.

56. Gross, B. L., Henk, A. D., Richards, C. M., Fazio, G., & Volk, G. M. (2014). Genetic diversity in Malus× domestica (Rosaceae) through time in response to domestication. *American journal of botany*, *101*(10), 1770-1779.

57. Guan, Y., Peace, C., Rudell, D., Verma, S., & Evans, K. (2015). QTLs detected for individual sugars and soluble solids content in apple. *Molecular Breeding*, *35*(6), 1-13.

58. Han, Y., & Korban, S. S. (2021). Genetic and physical mapping of the apple genome. In *The apple genome* (pp. 131-168). Springer, Cham.

59. Harada, T., Kurahashi, W., Yanai, M., Wakasa, Y., & Satoh, T. (2005). Involvement of cell proliferation and cell enlargement in increasing the fruit size of Malus species. *Scientia horticulturae*, *105*(4), 447-456.

60. Harris, S. A., Robinson, J. P., & Juniper, B. E. (2002). Genetic clues to the origin of the apple. *TRENDS in Genetics*, *18*(8), 426-430.

61. Hu, C. G., Hao, Y. J., Honda, C., Kita, M., & Moriguchi, T. (2003). Putative PIP1 genes isolated from apple: expression analyses during fruit development and under osmotic stress. *Journal of Experimental Botany*, *54*(390), 2193-2194.

62. Huang, M., Liu, X., Zhou, Y., Summers, R. M., & Zhang, Z. (2019). BLINK: a package for the next level of genome-wide association studies with both individuals and markers in the millions. *Gigascience*, *8*(2), giy154.

63. Iezzoni, A., Peace, C., Main, D., Bassil, N., Coe, M., Finn, C., ... & Yue, C. (2015, June). RosBREED2: progress and future plans to enable DNA-informed breeding in the Rosaceae. In *XIV EUCARPIA Symposium on Fruit Breeding and Genetics 1172* (pp. 115-118).

64. Jaeger, S. R., Antúnez, L., Ares, G., Swaney-Stueve, M., Jin, D., & Harker, F. R. (2018). Quality perceptions regarding external appearance of apples: Insights from experts and consumers in four countries. *Postharvest biology and technology*, *146*, 99-107.

65. Jagodzik, P., Tajdel-Zielinska, M., Ciesla, A., Marczak, M., & Ludwikow, A. (2018). Mitogen-activated protein kinase cascades in plant hormone signaling. *Frontiers in plant science*, *9*, 1387.

66. Janick, J., & Moore, J. N. (Eds.). (1996). *Fruit breeding, tree and tropical fruits* (Vol. 1). John Wiley & Sons.

67. Janssen, B. J., Thodey, K., Schaffer, R. J., Alba, R., Balakrishnan, L., Bishop, R., ... & Ward, S. (2008). Global gene expression analysis of apple fruit development from the floral bud to ripe fruit. *BMC Plant Biology*, *8*(1), 1-29.

68. Johnston, J. W., Gunaseelan, K., Pidakala, P., Wang, M., & Schaffer, R. J. (2009). Co-ordination of early and late ripening events in apples is regulated through differential sensitivities to ethylene. *Journal of experimental botany*, *60*(9), 2689-2699.

69. Jung, M., Keller, B., Roth, M., Aranzana, M. J., Auwerkerken, A., Guerra, W., ... & Patocchi, A. (2022). Genetic architecture and genomic predictive ability of apple quantitative traits across environments. *Horticulture research*, *9*.

70. Jung, M., Roth, M., Aranzana, M. J., Auwerkerken, A., Bink, M., Denancé, C., ... & Muranty, H. (2020). The apple REFPOP—a reference population for genomics-assisted breeding in apple. *Horticulture research*, *7*.

71. Jung, S., Lee, T., Cheng, C. H., Buble, K., Zheng, P., Yu, J., ... & Main, D. (2019). 15 years of GDR: New data and functionality in the Genome Database for Rosaceae. *Nucleic acids research*, *47*(D1), D1137-D1145.

72. Kajitani, R., Toshimoto, K., Noguchi, H., Toyoda, A., Ogura, Y., Okuno, M., ... & Itoh, T. (2014). Efficient de novo assembly of highly heterozygous genomes from whole-genome shotgun short reads. *Genome research*, *24*(8), 1384-1395.

73. Kenis, K., Keulemans, J., & Davey, M. W. (2008). Identification and stability of QTLs for fruit quality traits in apple. *Tree Genetics & Genomes*, *4*(4), 647-661.

74. Khanal, B. P., & Knoche, M. (2014). Mechanical properties of apple skin are determined by epidermis and hypodermis. *Journal of the American Society for Horticultural Science*, *139*(2), 139-147.

75. Kilara, A., & Buren, J. P. V. (1989). Clarification of apple juice. In *Processed apple products* (pp. 83-96). Springer, New York, NY.

76. Korban, S. S., & Skirvin, R. M. (1984). Nomenclature of the cultivated apple. *HortScience*, *19*(2), 177-180.

77. Kotoda, N., Wada, M., Komori, S., Kidou, S. I., Abe, K., Masuda, T., & Soejima, J. (2000). Expression pattern of homologues of floral meristem identity genes LFY and AP1 during flower development in apple. *Journal of the American Society for Horticultural Science*, *125*(4), 398-403.

78. Koutinas, N., Pepelyankov, G., & Lichev, V. (2010). Flower induction and flower bud development in apple and sweet cherry. *Biotechnology & Biotechnological Equipment*, *24*(1), 1549-1558.

79. Kumar, R., Khurana, A., & Sharma, A. K. (2013). Role of plant hormones and their interplay in development and ripening of fleshy fruits. *Journal of experimental botany*, *65*(16), 4561-4575.

80. Kumar, S., Rowan, D., Hunt, M., Chagné, D., Whitworth, C., & Souleyre, E. (2015). Genome-wide scans reveal genetic architecture of apple flavour volatiles. *Molecular breeding*, *35*(5), 1-16.

81. Kumar, S., Volz, R. K., Alspach, P. A., & Bus, V. G. (2010). Development of a recurrent apple breeding programme in New Zealand: a synthesis of results, and a proposed revised breeding strategy. *Euphytica*, *173*(2), 207-222.

82. La Belle, R. L. (1981). Apple quality characteristics as related to various processed products.

83. Larsen, B., Migicovsky, Z., Jeppesen, A. A., Gardner, K. M., Toldam-Andersen, T. B., Myles, S., ... & Pedersen, C. (2019). Genome-wide association studies in apple reveal loci for aroma volatiles, sugar composition, and harvest date. *The Plant Genome*, *12*(2), 180104.

84. Larsen, B., Migicovsky, Z., Jeppesen, A. A., Gardner, K. M., Toldam-Andersen, T. B., Myles, S., ... & Pedersen, C. (2019). Genome-wide association studies in apple reveal loci for aroma volatiles, sugar composition, and harvest date. *The Plant Genome*, *12*(2), 180104.

85. Larsen, B., Ørgaard, M., Toldam-Andersen, T. B., & Pedersen, C. (2016). A high-throughput method for genotyping S-RNase alleles in apple. *Molecular Breeding*, *36*(3), 1-10.

86. Laurens, F., (2010). Final Report Summary - FRUIT BREEDOMICS (Integrated approach for increasing breeding efficiency in fruit tree crops), Available: at https://cordis.europa.eu/project/id/265582/reporting/es

87. Lauri, P. É. (2016, October). Apple tree architecture and cultivation-a tree in a system. In *I International Apple Symposium 1261* (pp. 173-184).

88. Lauri, P. É., & Laurens, F. (2005). Architectural types in apple (Malus X domestica Borkh.). *Crops: growth, quality and biotechnology*, 1300-1314.

89. Lespinasse, J.M., Delort, J.F., (1993). Regulation of fruiting in apple. Acta Horticulturae 349.

90. Li, S. (2015). The Arabidopsis thaliana TCP transcription factors: a broadening horizon beyond development. *Plant signaling & behavior*, *10*(7), e1044192.

91. Li, T., Tan, D., Yang, X., & Wang, A. (2013). Exploring the apple genome reveals six ACC synthase genes expressed during fruit ripening. *Scientia Horticulturae*, *157*, 119-123.

92. Li, T., Xu, Y., Zhang, L., Ji, Y., Tan, D., Yuan, H., & Wang, A. (2017). The jasmonate-activated transcription factor MdMYC2 regulates ETHYLENE RESPONSE FACTOR and ethylene biosynthetic genes to promote ethylene biosynthesis during apple fruit ripening. *The Plant Cell*, *29*(6), 1316-1334.

93. Li, X., Kui, L., Zhang, J., Xie, Y., Wang, L., Yan, Y., ... & Guan, Q. (2016). Improved hybrid de novo genome assembly of domesticated apple (Malus x domestica). *Gigascience*, *5*(1), s13742-016.

94. Li, Z., Wang, L., He, J., Li, X., Hou, N., Guo, J., ... & Guan, Q. (2022). Chromosome-scale reference genome provides insights into the genetic origin and grafting-mediated stress tolerance of Malus prunifolia. *Plant Biotechnology Journal*.

95. Lian, Q., Fu, Q., Xu, Y., Hu, Z., Zheng, J., Zhang, A., ... & Wang, H. (2021). QTLs and candidate genes analyses for fruit size under domestication and differentiation in melon (Cucumis melo L.) based on high resolution maps. *BMC plant biology*, *21*(1), 1-13.

96. Licausi, F., Ohme-Takagi, M., & Perata, P. (2013). APETALA 2/Ethylene Responsive Factor (AP 2/ERF) transcription factors: Mediators of stress responses and developmental programs. *New Phytologist*, *199*(3), 639-649.

97. Liebhard, R., Kellerhals, M., Pfammatter, W., Jertmini, M., & Gessler, C. (2003). Mapping quantitative physiological traits in apple (Malus× domestica Borkh.). *Plant molecular biology*, *52*(3), 511-526.

98. Liu, L., Wang, Z., Liu, J., Liu, F., Zhai, R., Zhu, C., ... & Xu, L. (2018). Histological, hormonal and transcriptomic reveal the changes upon gibberellin-induced parthenocarpy in pear fruit. *Horticulture research*, *5*.

99. Liu, M., Gomes, B. L., Mila, I., Purgatto, E., Peres, L. E., Frasse, P., ... & Pirrello, J. (2016). Comprehensive profiling of ethylene response factor expression identifies ripening-associated ERF genes and their link to key regulators of fruit ripening in tomato. *Plant Physiology*, *170*(3), 1732-1744.

100. Liu, X., Huang, M., Fan, B., Buckler, E. S., & Zhang, Z. (2016). Iterative usage of fixed and random effect models for powerful and efficient genome-wide association studies. *PLoS genetics*, *12*(2), e1005767.

101. Ma, B., Zhao, S., Wu, B., Wang, D., Peng, Q., Owiti, A., ... & Han, Y. (2016). Construction of a high density linkage map and its application in the identification of QTLs for soluble sugar and organic acid components in apple. *Tree genetics & genomes*, *12*(1), 1-10.

102. Ma, H., & DePamphilis, C. (2000). The ABCs of floral evolution. *Cell*, *101*(1), 5-8.

103. MacDaniels, L. H., & Heinicke, A. J. (1929). Pollination and other factors affecting the set of fruit with special reference to the apple. *Bull. Cornell Agric. Exp. Stn.*

104. Malladi, A. (2020). Molecular physiology of fruit growth in apple. *Horticultural reviews*, *47*, 1-42.

105. Malladi, A., & Johnson, L. K. (2011). Expression profiling of cell cycle genes reveals key facilitators of cell production during carpel development, fruit set, and fruit growth in apple (Malus× domestica Borkh.). *Journal of experimental botany*, *62*(1), 205-219.

106. Malnoy, M., Viola, R., Jung, M. H., Koo, O. J., Kim, S., Kim, J. S., ... & Nagamangala Kanchiswamy, C. (2016). DNA-free genetically edited grapevine and apple protoplast using CRISPR/Cas9 ribonucleoproteins. *Frontiers in plant science*, *7*, 1904.

107. Marini, R. P., & Fazio, G. (2018). Apple rootstocks: History, physiology, management, and breeding. *Horticultural Reviews*, *45*, 197-312.

108. Markussen, T., Krüger, J., Schmidt, H., & Dunemann, F. (1995). Identification of PCR-based markers linked to the powdery-mildew-resistance gene Pl1 from Malus robusta in cultivated apple. *Plant Breeding*, *114*(6), 530-534.

109. Marquard, R. D., & Chan, C. R. (1995). Identifying Crabapple cultivars by isozymes. *Journal of the American Society for Horticultural Science*, *120*(5), 706-709.

110. Martínez-Martínez, C., Gonzalo, M. J., Sipowicz, P., Campos, M., Martínez-Fernández, I., Leida, C., ... & Monforte, A. J. (2022). A cryptic variation in a member of the Ovate Family Proteins is underlying the melon fruit shape QTL fsqs8. 1. *Theoretical and Applied Genetics*, *135*(3), 785-801.

111. Marwal, A., & Gaur, R. K. (2020). Molecular markers: tool for genetic analysis. In *Animal Biotechnology* (pp. 353-372). Academic Press.

112. Matsumoto, S., Hoshi, N., Tsuchiya, T., Soejima, J., Komori, S., & Ejiri, S. (1994). S-RNase like genomic sequences in apple for DNA fingerprinting. *Genetic Improvement of Horticultural Crops by Biotechnology 392*, 265-274.

113. McClure, K. A., Gardner, K. M., Douglas, G. M., Song, J., Forney, C. F., DeLong, J., ... & Myles, S. (2018). A genome-wide association study of apple quality and scab resistance. *The Plant Genome*, *11*(1), 170075.

114. McClure, K. A., Gong, Y., Song, J., Vinqvist-Tymchuk, M., Campbell Palmer, L., Fan, L., ... & Myles, S. (2019). Genome-wide association studies in apple reveal loci of large effect controlling apple polyphenols. *Horticulture research*, *6*.

115. Meshram, V., Patil, K., Meshram, V., Hanchate, D., & Ramkteke, S. D. (2021). Machine learning in agriculture domain: A state-of-art survey. *Artificial Intelligence in the Life Sciences*, *1*, 100010.

116. Migicovsky, Z., Gardner, K. M., Money, D., Sawler, J., Bloom, J. S., Moffett, P., ... & Myles, S. (2016). Genome to phenome mapping in apple using historical data. *The Plant Genome*, *9*(2), plantgenome2015-11.

117. Moradi, E., Abdolshahnejad, M., Hassangavyar, M. B., Ghoohestani, G., da Silva, A. M., Khosravi, H., & Cerdà, A. (2021). Machine learning approach to predict susceptible growth regions of Moringa peregrina (Forssk). *Ecological Informatics*, *62*, 101267.

118. Morgan, D. R., Soltis, D. E., & Robertson, K. R. (1994). Systematic and evolutionary implications of rbcL sequence variation in Rosaceae. *American Journal of Botany*, *81*(7), 890-903.

119. Mouad, Z., Helin, D., Rasti, P., Maria-Jose Aranzana., Dujak, C., Rousseau D. Toward objective variety testing score based on computer vision and unsupervised machine learning Application to Apple Shape. In press. Biosystems Engineering

120. Nakagawa, S., Bukovac, M. J., Hirata, N., & Kurooka, H. (1968). Morphological studies of gibberellin-induced parthenocarpic and asymmetric growth in apple and Japanese pear fruits. *Journal of the Japanese Society for Horticultural Science*, *37*(1), 9-19.

121. Nankar, A. N., Tringovska, I., Grozeva, S., Todorova, V., & Kostova, D. (2020). Application of high-throughput phenotyping tool Tomato Analyzer to characterize Balkan Capsicum fruit diversity. *Scientia Horticulturae*, *260*, 108862.

122. Nikiforova, S. V., Cavalieri, D., Velasco, R., & Goremykin, V. (2013). Phylogenetic analysis of 47 chloroplast genomes clarifies the contribution of wild species to the domesticated apple maternal line. *Molecular biology and evolution*, *30*(8), 1751-1760.

123. Nishihama, R., Ishikawa, M., Araki, S., Soyano, T., Asada, T., & Machida, Y. (2001). The NPK1 mitogen-activated protein kinase kinase kinase is a regulator of cell-plate formation in plant cytokinesis. *Genes & Development*, *15*(3), 352-363.

124. O'Rourke, J. A. (2014). Genetic and physical map correlation. *eLS*.

125. Peace, C. P., Bianco, L., Troggio, M., Van de Weg, E., Howard, N. P., Cornille, A., ... & Vanderzande, S. (2019). Apple whole genome sequences: recent advances and new prospects. *Horticulture Research*, *6*.

126. Pereira, L., Santo Domingo, M., Ruggieri, V., Argyris, J., Phillips, M. A., Zhao, G., ... & Garcia-Mas, J. (2020). Genetic dissection of climacteric fruit ripening in a melon population segregating for ripening behavior. *Horticulture research*, *7*.

127. Potts, S. M., Han, Y., Khan, M. A., Kushad, M. M., Rayburn, A. L., & Korban, S. S. (2012). Genetic diversity and characterization of a core collection of Malus germplasm using simple sequence repeats (SSRs). *Plant Molecular Biology Reporter*, *30*(4), 827-837.

128. Potts, S. M., Khan, M. A., Han, Y., Kushad, M. M., & Korban, S. S. (2014). Identification of quantitative trait loci (QTLs) for fruit quality traits in apple. *Plant molecular biology reporter*, *32*(1), 109-116.

129. Pratt, C. (1993). Apple Trees: Morphology and Anatomy, HORTICULTURAL REVIEWS, Vol. 12.

130. Pratt, C. (2011). Apple flower and fruit: morphology and anatomy. *Horticultural Reviews*, *10*, 273-308.

131. Qi, Y. (2012). Random forest for bioinformatics. In *Ensemble machine learning* (pp. 307-323). Springer, Boston, MA.

132. Qian, G. Z., Liu, L. F., & Tang, G. G. (2006, January). A new section in Malus (Rosaceae) from China. In *Annales Botanici Fennici* (pp. 68-73). Finnish Zoological and Botanical Publishing Board.

133. Robinson, J. P., Harris, S. A., & Juniper, B. E. (2001). Taxonomy of the genus Malus Mill.(Rosaceae) with emphasis on the cultivated apple, Malus domestica Borkh. *Plant Systematics and Evolution*, *226*(1), 35-58.

134. Rodríguez, G. R., Muños, S., Anderson, C., Sim, S. C., Michel, A., Causse, M., ... & van Der Knaap, E. (2011). Distribution of SUN, OVATE, LC, and FAS in the tomato germplasm and the relationship to fruit shape diversity. *Plant physiology*, *156*(1), 275-285.

135. Rymenants, M., van de Weg, E., Auwerkerken, A., De Wit, I., Czech, A., Nijland, B., ... & Keulemans, W. (2020). Detection of QTL for apple fruit acidity and sweetness using sensorial evaluation in multiple pedigreed full-sib families. *Tree Genetics & Genomes*, *16*(5), 1-16.

136. Rymenants, M., van de Weg, E., Auwerkerken, A., De Wit, I., Czech, A., Nijland, B., ... & Keulemans, W. (2020). Detection of QTL for apple fruit acidity and sweetness using sensorial evaluation in multiple pedigreed full-sib families. *Tree Genetics & Genomes*, *16*(5), 1-16.

137. Sampedro, J., & Cosgrove, D. J. (2005). The expansin superfamily. *Genome biology*, *6*(12), 1-11.

138. Sánchez-Galán, J. E., Barranco, F. R., Reyes, J. S., Quirós-McIntire, E. I., Jiménez, J. U., & Fábrega, J. R. (2019). Using Supervised Classification Methods for the Analysis of Multi-spectral Signatures of Rice Varieties in Panama.

139. Schmidt, H. (1994). Progress in combining mildew resistance from Malus robusta and Malus zumi with fruit quality. In *Progress in Temperate Fruit Breeding* (pp. 3-6). Springer, Dordrecht.

140. Schneider, D., Stern, R. A., & Goldway, M. (2005). A comparison between semi-and fully compatible apple pollinators grown under suboptimal pollination conditions. *HortScience*, *40*(5), 1280-1282.

141. Segura, V., Vilhjálmsson, B. J., Platt, A., Korte, A., Seren, Ü., Long, Q., & Nordborg, M. (2012). An efficient multi-locus mixed-model approach for genome-wide association studies in structured populations. *Nature genetics*, *44*(7), 825-830.

142. Serrani, J. C., Fos, M., Atarés, A., & García-Martínez, J. L. (2007). Effect of gibberellin and auxin on parthenocarpic fruit growth induction in the cv Micro-Tom of tomato. *Journal of Plant Growth Regulation*, *26*(3), 211-221.

143. Sierra-Orozco, E., Shekasteband, R., Illa-Berenguer, E., Snouffer, A., van der Knaap, E., Lee, T. G., & Hutton, S. F. (2021). Identification and characterization of GLOBE, a major gene controlling fruit shape and impacting fruit size and marketability in tomato. *Horticulture research*, *8*.

144. Spengler, R. N. (2019). Origins of the apple: the role of megafaunal mutualism in the domestication of Malus and rosaceous trees. *Frontiers in plant science*, *10*, 617.

145. Srivastava, A., & Handa, A. K. (2005). Hormonal regulation of tomato fruit development: a molecular perspective. *Journal of plant growth regulation*, *24*(2), 67-82.

146. Štampar, F., & Smole, J. (1990). *Identification of apple and sweet cherry cultivars, peach hybrids and apricot ecotypes by isozyme phenotyping* (pp. 155-162).

147. Sun, X., Jiao, C., Schwaninger, H., Chao, C. T., Ma, Y., Duan, N., ... & Fei, Z. (2020). Phased diploid genome assemblies and pan-genomes provide insights into the genetic history of apple domestication. *Nature genetics*, *52*(12), 1423-1432.

148. Szalatnay, D., & Bauermeister, R. (2006). Obst-Deskriptoren NAP. *Stutz Druck AG*, *8820*.

149. Tam, V., Patel, N., Turcotte, M., Bossé, Y., Paré, G., & Meyre, D. (2019). Benefits and limitations of genome-wide association studies. *Nature Reviews Genetics*, *20*(8), 467-484.

150. Tanksley, S. D. (2004). The genetic, developmental, and molecular bases of fruit size and shape variation in tomato. *The plant cell*, *16*(suppl_1), S181-S189.

151. Teh, S. L., Kostick, S. A., & Evans, K. M. (2021). Genetics and breeding of apple scions. In *The apple genome* (pp. 73-103). Springer, Cham.

152. Troggio, M. I. C. H. E. L. A., Gleave, A., Salvi, S., Chagné, D., Cestaro, A., Kumar, S., ... & Gardiner, S. E. (2012). Apple, from genome to breeding. *Tree genetics & genomes*, *8*(3), 509-529.

153. Urrestarazu, J., Denancé, C., Ravon, E., Guyader, A., Guisnel, R., Feugey, L., ... & Durel, C. E. (2016). Analysis of the genetic diversity and structure across a wide range of germplasm reveals prominent gene flow in apple at the European level. *BMC plant biology*, *16*(1), 1-20.

154. Urrestarazu, J., Muranty, H., Denancé, C., Leforestier, D., Ravon, E., Guyader, A., ... & Durel, C. E. (2017). Genome-wide association mapping of flowering and ripening periods in apple. *Frontiers in plant science*, *8*, 1923.

155. Velasco, R., Zharkikh, A., Affourtit, J., Dhingra, A., Cestaro, A., Kalyanaraman, A., ... & Viola, R. (2010). The genome of the domesticated apple (Malus× domestica Borkh.). *Nature genetics*, *42*(10), 833-839.

156. Verma, S., Evans, K., Guan, Y., Luby, J. J., Rosyara, U. R., Howard, N. P., ... & Peace, C. P. (2019). Two large-effect QTLs, Ma and Ma3, determine genetic potential for acidity in apple fruit: breeding insights from a multi-family study. *Tree Genetics & Genomes*, *15*(2), 1-17.

157. Wang, H., Jones, B., Li, Z., Frasse, P., Delalande, C., Regad, F., ... & Bouzayen, M. (2005). The tomato Aux/IAA transcription factor IAA9 is involved in fruit development and leaf morphogenesis. *The Plant Cell*, *17*(10), 2676-2692.

158. WAPA. (2019). THE WORLD APPLE AND PEAR ASSOCIATION. Available at:http://www.wapaassociation.org/docs/2019/European_summary_reduced.pdf

159. Watanabe, M., Segawa, H., Murakami, M., Sagawa, S., & Komori, S. (2008). Effects of plant growth regulators on fruit set and fruit shape of parthenocarpic apple fruits. *Journal of the Japanese society for Horticultural Science*, *77*(4), 350-357.

160. Watson, B. (2013). *Cider, hard and sweet: History, traditions, and making your own*. The Countryman Press.

161. Way, R. D., & McLellan, M. R. (1989). Apple cultivars for processing. In *Processed apple products* (pp. 1-29). Springer, New York, NY.

162. Wiersma, P. A., Zhang, H., Lu, C., Quail, A., & Toivonen, P. M. (2007). Survey of the expression of genes for ethylene synthesis and perception during maturation and ripening of 'Sunrise'and 'Golden Delicious' apple fruit. *Postharvest Biology and Technology*, *44*(3), 204-211.

163. Williams, R. R.; Maier, Maria (1977). Pseudocompatibility After Self-Pollination of the Apple Cox's Orange Pippin. Journal of Horticultural Science, 52(4), 475–483.

164. Xiang, Y., Huang, C. H., Hu, Y., Wen, J., Li, S., Yi, T., ... & Ma, H. (2017). Evolution of Rosaceae fruit types based on nuclear phylogeny in the context of geological times and genome duplication. *Molecular biology and evolution*, *34*(2), 262-281.

165. Yang, X., Song, J., Campbell-Palmer, L., Fillmore, S., & Zhang, Z. (2013). Effect of ethylene and 1-MCP on expression of genes involved in ethylene biosynthesis and perception during ripening of apple fruit. *Postharvest Biology and Technology*, *78*, 55-66.

166. Yao, J. L., Dong, Y. H., & Morris, B. A. (2001). Parthenocarpic apple fruit production conferred by transposon insertion mutations in a MADS-box transcription factor. *Proceedings of the National Academy of Sciences*, *98*(3), 1306-1311.

167. Yao, J. L., Dong, Y. H., Kvarnheden, A., & Morris, B. (1999). Seven MADS-box genes in apple are expressed in different parts of the fruit. *Journal of the American Society for Horticultural Science*, *124*(1), 8-13.

168. Yao, J. L., Xu, J., Cornille, A., Tomes, S., Karunairetnam, S., Luo, Z., ... & Gleave, A. P. (2015). A micro RNA allele that emerged prior to apple domestication may underlie fruit size evolution. *The Plant Journal*, *84*(2), 417-427.

169. Yao, J. L., Xu, J., Tomes, S., Cui, W., Luo, Z., Deng, C., ... & Gleave, A. P. (2018). Ectopic expression of the PISTILLATA homologous MdPI inhibits fruit tissue growth and changes fruit shape in apple. Plant Direct 2018: 1–11.

170. Yue, P., Lu, Q., Liu, Z., Lv, T., Li, X., Bu, H., ... & Wang, A. (2020). Auxin-activated MdARF5 induces the expression of ethylene biosynthetic genes to initiate apple fruit ripening. *New Phytologist*, *226*(6), 1781-1795.

171. Zhang, J., Zhang, W., Xiong, S., Song, Z., Tian, W., Shi, L., & Ma, X. (2021). Comparison of new hyperspectral index and machine learning models for prediction of winter wheat leaf water content. *Plant Methods*, *17*(1), 1-14.

172. Zhang, L., Hu, J., Han, X., Li, J., Gao, Y., Richards, C. M., ... & Cong, P. (2019). A high-quality apple genome assembly reveals the association of a retrotransposon and red fruit colour. *Nature communications*, *10*(1), 1-13.

173. Zhang, S., Xu, R., Gao, Z., Chen, C., Jiang, Z., & Shu, H. (2014). A genome-wide analysis of the expansin genes in Malus× Domestica. *Molecular genetics and genomics*, *289*(2), 225-236.

174. Zhang, Z., Ersoz, E., Lai, C. Q., Todhunter, R. J., Tiwari, H. K., Gore, M. A., ... & Buckler, E. S. (2010). Mixed linear model approach adapted for genome-wide association studies. *Nature genetics*, *42*(4), 355-360.

175. Zhou, H., Ma, R., Gao, L., Zhang, J., Zhang, A., Zhang, X., ... & Han, Y. (2021). A 1.7-Mb chromosomal inversion downstream of a PpOFP1 gene is responsible for flat fruit shape in peach. *Plant biotechnology journal*, *19*(1), 192-205.